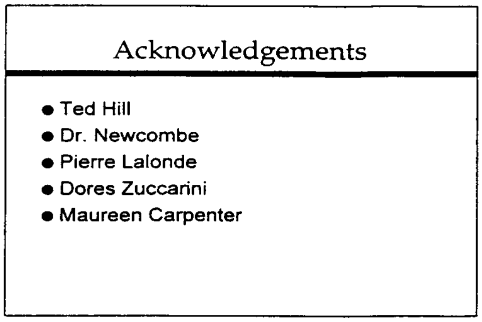
Chapter 12

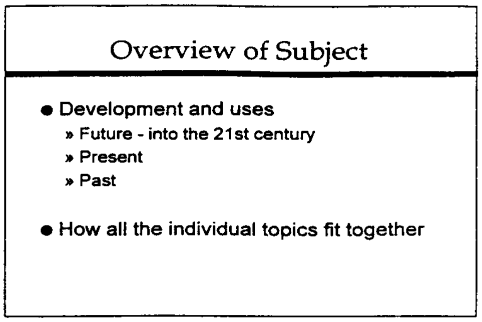
Tutorial on Record Linkage Slides Presentation

Martha E. Fair and Patricia Whitridge, Statistics Canada



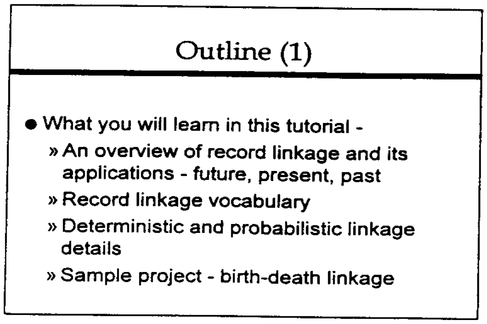
Acknowledgements

- Ted Hill
- Dr. Newcombe
- Pierre Lalonde
- Dores Zuccarini
- Maureen Carpenter



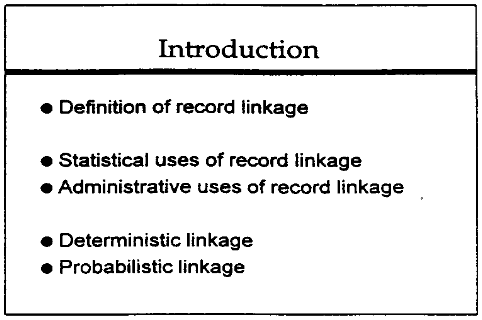
Overview of Subject

- Development and uses
  - » Future - into the 21st century
  - » Present
  - » Past
- How all the individual topics fit together



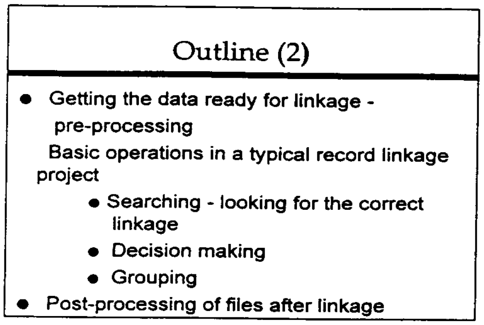
Outline (1)

- What you will learn in this tutorial -
  - » An overview of record linkage and its applications - future, present, past
  - » Record linkage vocabulary
  - » Deterministic and probabilistic linkage details
  - » Sample project - birth-death linkage



Introduction

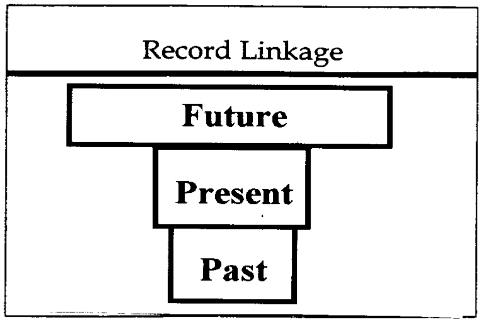
- Definition of record linkage
- Statistical uses of record linkage
- Administrative uses of record linkage
- Deterministic linkage
- Probabilistic linkage



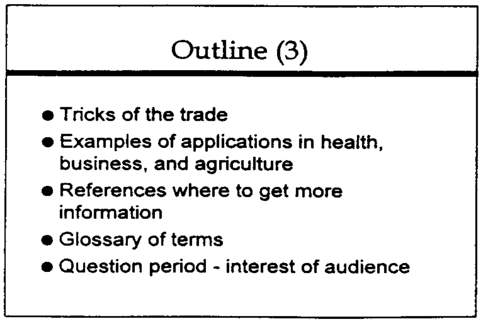
Outline (2)

- Getting the data ready for linkage - pre-processing

  Basic operations in a typical record linkage project
  - Searching - looking for the correct linkage
  - Decision making
  - Grouping
- Post-processing of files after linkage

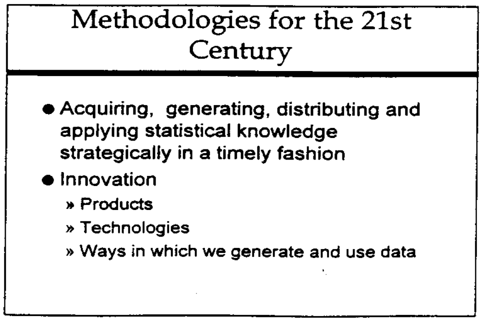


Record Linkage



Outline (3)

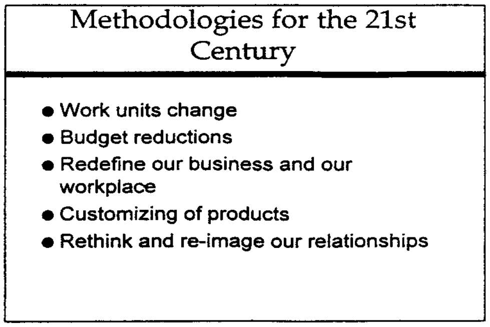
- Tricks of the trade
- Examples of applications in health, business, and agriculture
- References where to get more information
- Glossary of terms
- Question period - interest of audience



Methodologies for the 21st Century

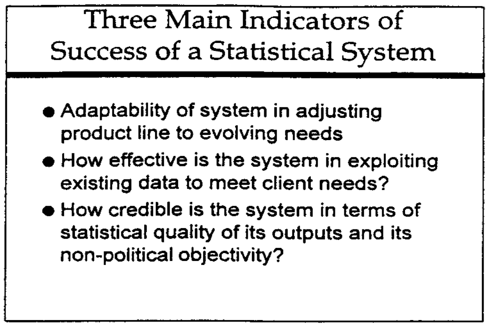
- Acquiring, generating, distributing and applying statistical knowledge strategically in a timely fashion
- Innovation
  - » Products
  - » Technologies
  - » Ways in which we generate and use data

Slides Presentation (cont'd)



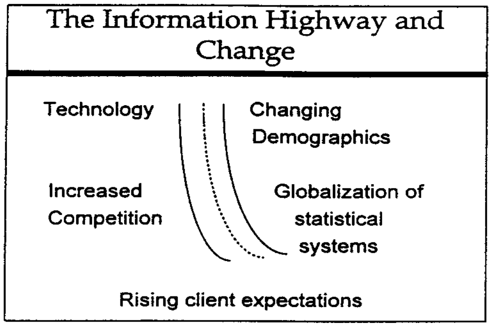
Methodologies for the 21st Century

- Work units change
- Budget reductions
- Redefine our business and our workplace
- Customizing of products
- Rethink and re-image our relationships



Three Main Indicators of Success of a Statistical System

- Adaptability of system in adjusting product line to evolving needs
- How effective is the system in exploiting existing data to meet client needs?
- How credible is the system in terms of statistical quality of its outputs and its non-political objectivity?



The Information Highway and Change

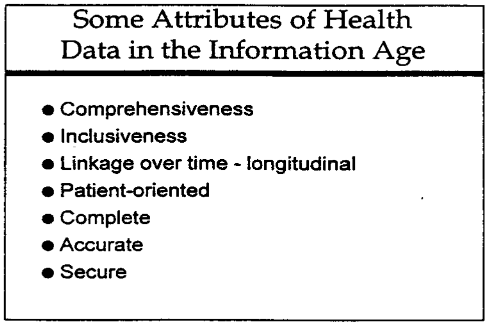
Technology

Changing Demographics

Increased Competition

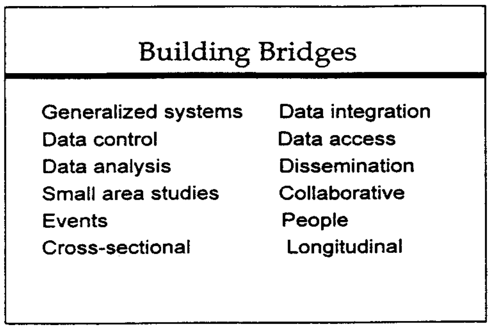
Globalization of statistical systems

Rising client expectations



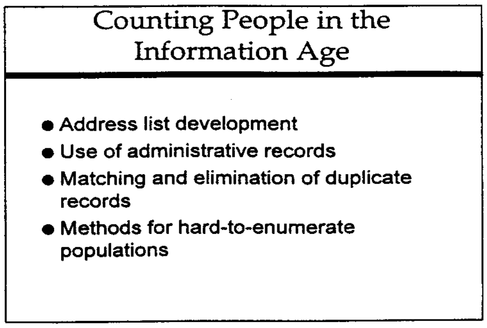
Some Attributes of Health Data in the Information Age

- Comprehensiveness
- Inclusiveness
- Linkage over time - longitudinal
- Patient-oriented
- Complete
- Accurate
- Secure



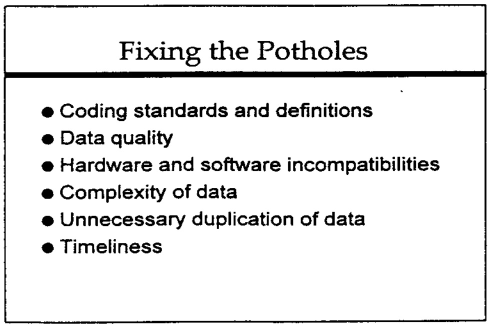
Building Bridges

Generalized systems	Data integration
Data control	Data access
Data analysis	Dissemination
Small area studies	Collaborative
Events	People
Cross-sectional	Longitudinal



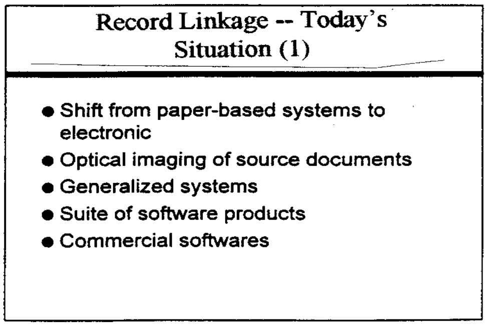
Counting People in the Information Age

- Address list development
- Use of administrative records
- Matching and elimination of duplicate records
- Methods for hard-to-enumerate populations



Fixing the Potholes

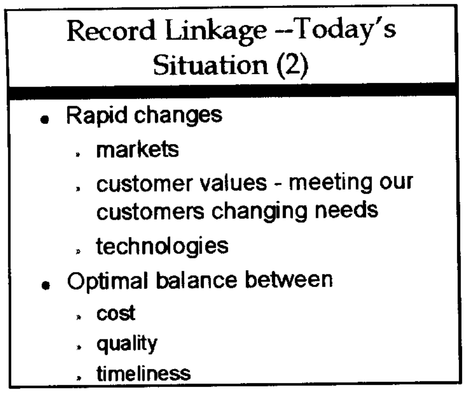
- Coding standards and definitions
- Data quality
- Hardware and software incompatibilities
- Complexity of data
- Unnecessary duplication of data
- Timeliness



Record Linkage -- Today's Situation (1)

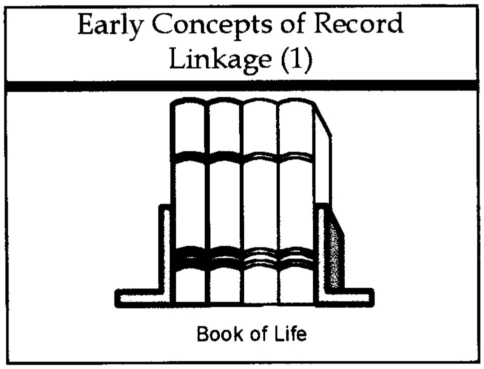
- Shift from paper-based systems to electronic
- Optical imaging of source documents
- Generalized systems
- Suite of software products
- Commercial softwares

Slides Presentation (cont'd)



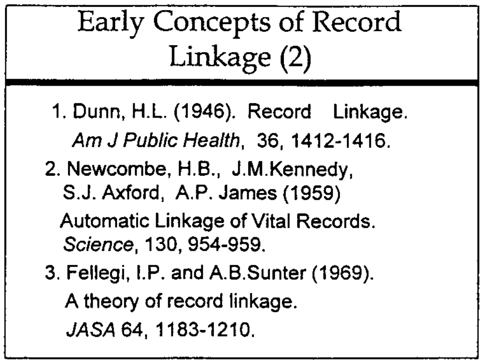
Record Linkage --Today's Situation (2)

- Rapid changes
  - markets
  - customer values - meeting our customers changing needs
  - technologies
- Optimal balance between
  - cost
  - quality
  - timeliness



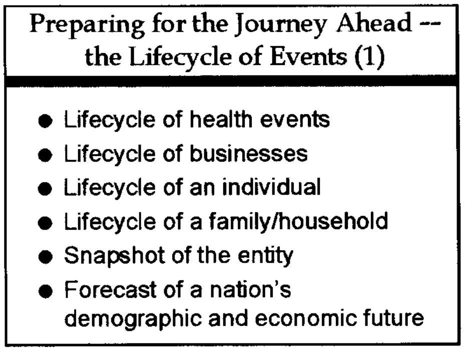
Early Concepts of Record Linkage (1)

Book of Life



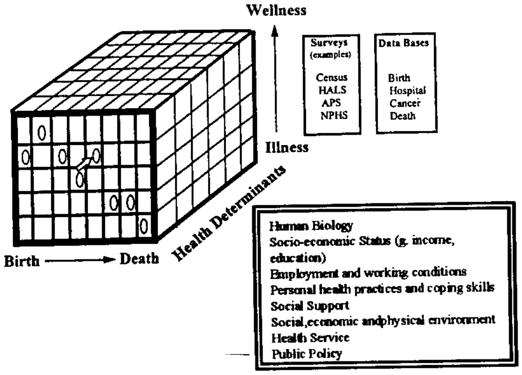
Early Concepts of Record Linkage (2)

1. Dunn, H.L. (1946). Record Linkage. *Am J Public Health*, 36, 1412-1416.
2. Newcombe, H.B., J.M.Kennedy, S.J. Axford, A.P. James (1959) Automatic Linkage of Vital Records. *Science*, 130, 954-959.
3. Fellegi, I.P. and A.B.Sunter (1969). A theory of record linkage. *JASA* 64, 1183-1210.



Preparing for the Journey Ahead -- the Lifecycle of Events (1)

- Lifecycle of health events
- Lifecycle of businesses
- Lifecycle of an individual
- Lifecycle of a family/household
- Snapshot of the entity
- Forecast of a nation's demographic and economic future



Wellness

Illness

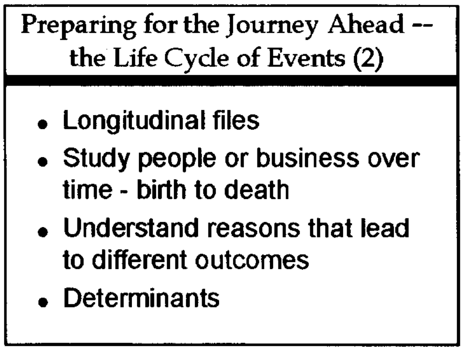
Birth → Death

Health Determinants

Surveys (examples)
Census
HALS
APS
NPHS

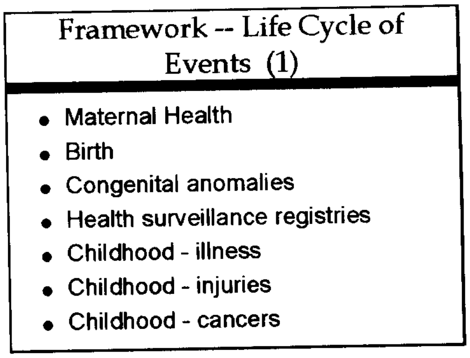
Data Bases
Birth
Hospital
Cancer
Death

Human Biology
Socio-economic Status (g. income, education)
Employment and working conditions
Personal health practices and coping skills
Social Support
Social,economic andphysical environment
Health Service
Public Policy



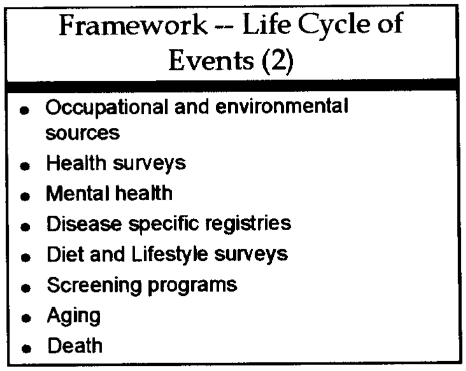
Preparing for the Journey Ahead -- the Life Cycle of Events (2)

- Longitudinal files
- Study people or business over time - birth to death
- Understand reasons that lead to different outcomes
- Determinants



Framework -- Life Cycle of Events (1)

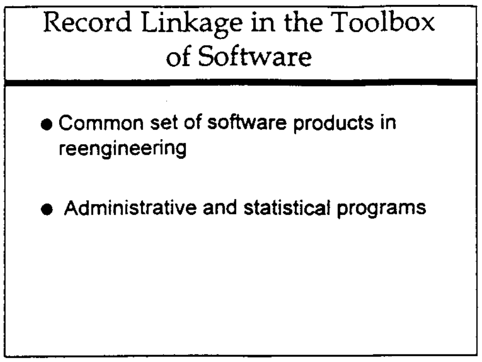
- Maternal Health
- Birth
- Congenital anomalies
- Health surveillance registries
- Childhood - illness
- Childhood - injuries
- Childhood - cancers



Framework -- Life Cycle of Events (2)

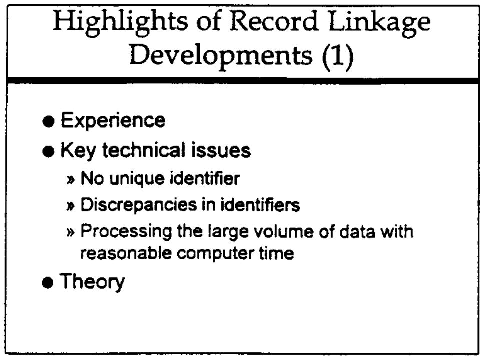
- Occupational and environmental sources
- Health surveys
- Mental health
- Disease specific registries
- Diet and Lifestyle surveys
- Screening programs
- Aging
- Death

Slides Presentation (cont'd)



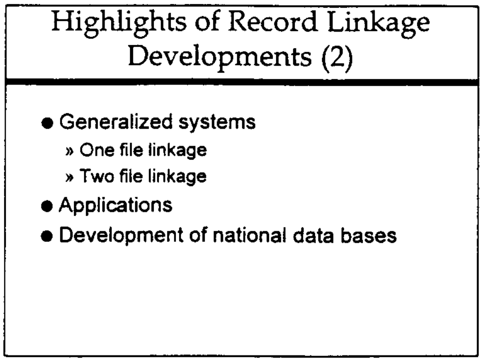
Record Linkage in the Toolbox of Software

- Common set of software products in reengineering
- Administrative and statistical programs



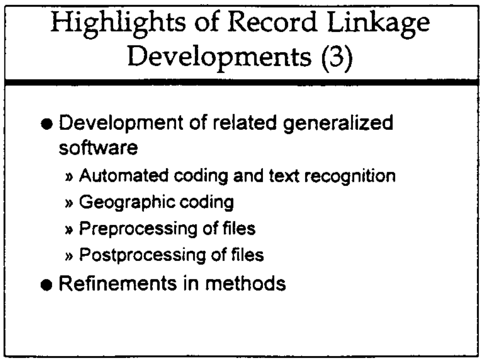
Highlights of Record Linkage Developments (1)

- Experience
- Key technical issues
  - » No unique identifier
  - » Discrepancies in identifiers
  - » Processing the large volume of data with reasonable computer time
- Theory



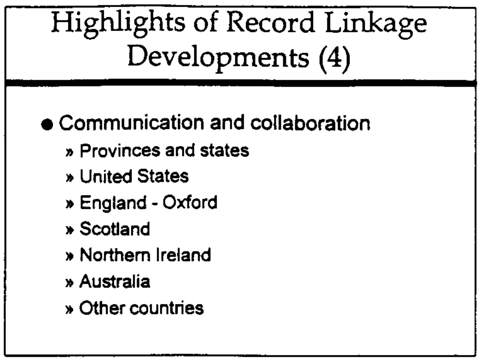
Highlights of Record Linkage Developments (2)

- Generalized systems
  - » One file linkage
  - » Two file linkage
- Applications
- Development of national data bases



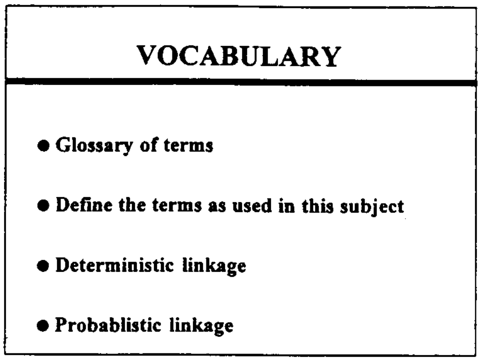
Highlights of Record Linkage Developments (3)

- Development of related generalized software
  - » Automated coding and text recognition
  - » Geographic coding
  - » Preprocessing of files
  - » Postprocessing of files
- Refinements in methods



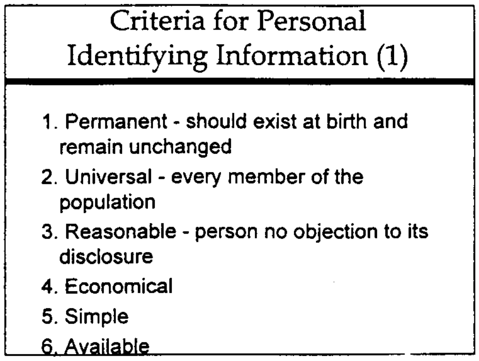
Highlights of Record Linkage Developments (4)

- Communication and collaboration
  - » Provinces and states
  - » United States
  - » England - Oxford
  - » Scotland
  - » Northern Ireland
  - » Australia
  - » Other countries



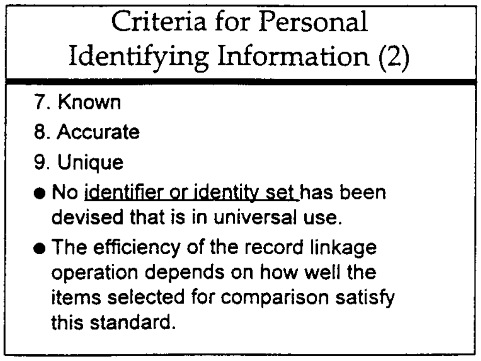
VOCABULARY

- **Glossary of terms**
- **Define the terms as used in this subject**
- **Deterministic linkage**
- **Probablistic linkage**



Criteria for Personal Identifying Information (1)

1. Permanent - should exist at birth and remain unchanged
2. Universal - every member of the population
3. Reasonable - person no objection to its disclosure
4. Economical
5. Simple
6. Available

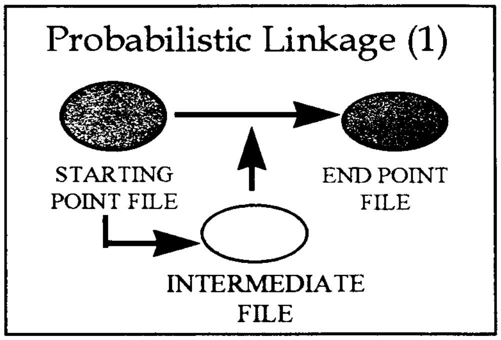


Criteria for Personal Identifying Information (2)

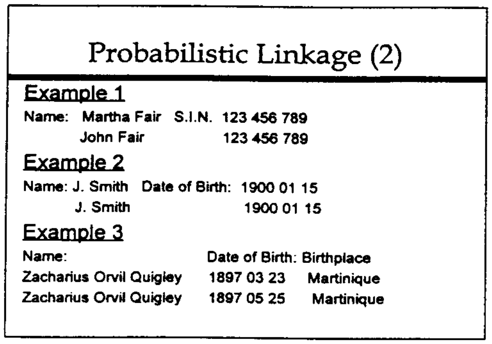
7. Known
8. Accurate
9. Unique

- No identifier or identity set has been devised that is in universal use.
- The efficiency of the record linkage operation depends on how well the items selected for comparison satisfy this standard.

Slides Presentation (cont'd)



Probabilistic Linkage (1)



Probabilistic Linkage (2)

Example 1

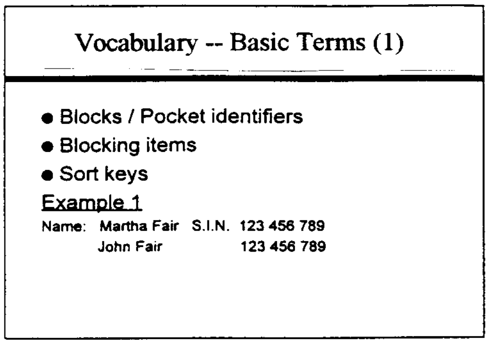
Name:	Martha Fair	S.I.N.	123 456 789
	John Fair		123 456 789

Example 2

Name:	J. Smith	Date of Birth:	1900 01 15
	J. Smith		1900 01 15

Example 3

Name:	Date of Birth:	Birthplace
Zacharius Orvil Quigley	1897 03 23	Martinique
Zacharius Orvil Quigley	1897 05 25	Martinique

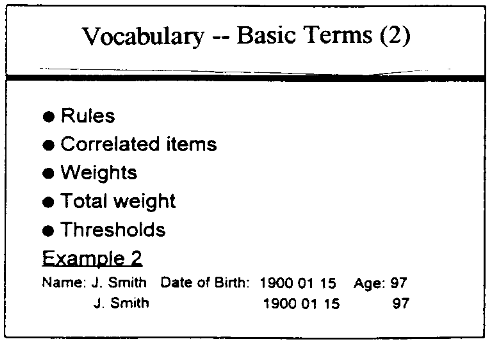


Vocabulary -- Basic Terms (1)

- Blocks / Pocket identifiers
- Blocking items
- Sort keys

Example 1

Name:	Martha Fair	S.I.N.	123 456 789
	John Fair		123 456 789

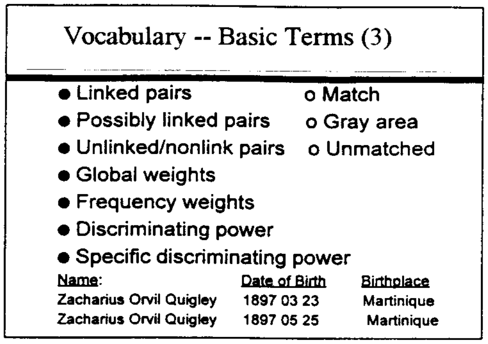


Vocabulary -- Basic Terms (2)

- Rules
- Correlated items
- Weights
- Total weight
- Thresholds

Example 2

Name:	J. Smith	Date of Birth:	1900 01 15	Age: 97
	J. Smith		1900 01 15	97

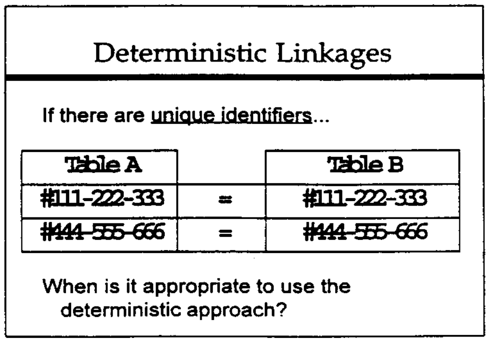


Vocabulary -- Basic Terms (3)

● Linked pairs	o Match
● Possibly linked pairs	o Gray area
● Unlinked/nonlink pairs	o Unmatched

- Global weights
- Frequency weights
- Discriminating power
- Specific discriminating power

Name:	Date of Birth	Birthplace
Zacharius Orvil Quigley	1897 03 23	Martinique
Zacharius Orvil Quigley	1897 05 25	Martinique

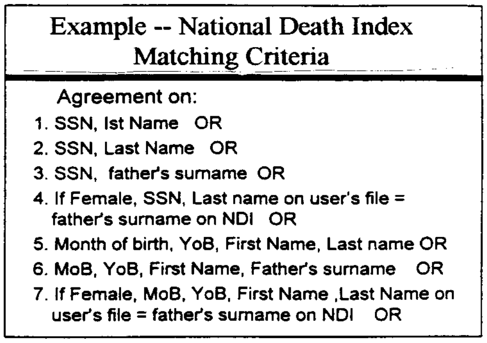


Deterministic Linkages

If there are unique identifiers...

Table A		Table B
#111-222-333	=	#111-222-333
~~#444-555-666~~	=	~~#444-555-666~~

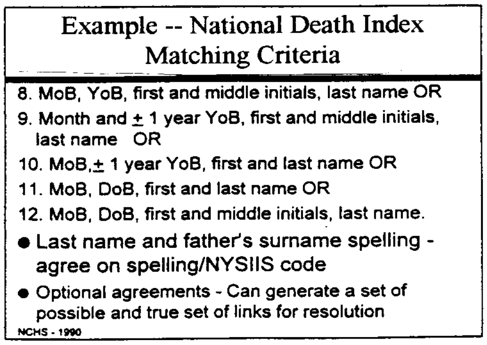
When is it appropriate to use the deterministic approach?



Example -- National Death Index Matching Criteria

Agreement on:

1. SSN, Ist Name OR
2. SSN, Last Name OR
3. SSN, father's surname OR
4. If Female, SSN, Last name on user's file = father's surname on NDI OR
5. Month of birth, YoB, First Name, Last name OR
6. MoB, YoB, First Name, Father's surname OR
7. If Female, MoB, YoB, First Name ,Last Name on user's file = father's surname on NDI OR



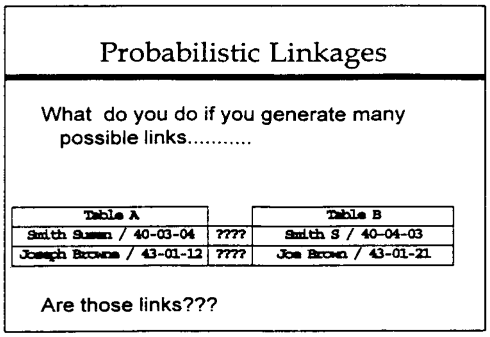
Example -- National Death Index Matching Criteria

8. MoB, YoB, first and middle initials, last name OR
9. Month and ± 1 year YoB, first and middle initials, last name OR
10. MoB,± 1 year YoB, first and last name OR
11. MoB, DoB, first and last name OR
12. MoB, DoB, first and middle initials, last name.

- Last name and father's surname spelling - agree on spelling/NYSIIS code
- Optional agreements - Can generate a set of possible and true set of links for resolution

NCHS - 1990

Slides Presentation (cont'd)

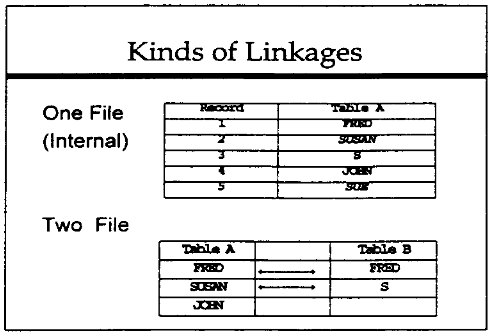


Probabilistic Linkages

What do you do if you generate many possible links..........

Table A		Table B
Smith Susan / 40-03-04	????	Smith S / 40-04-03
Joseph Browne / 43-01-12	????	Joe Brown / 43-01-21

Are those links???



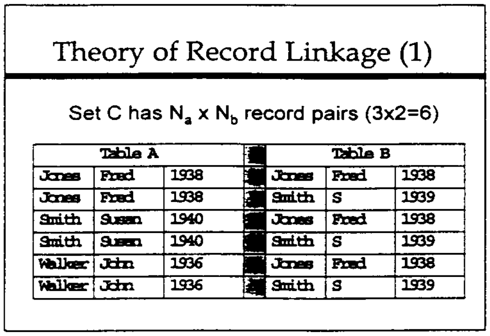
Kinds of Linkages

One File
(Internal)

Record	Table A
1	FRED
2	SUSAN
3	S
4	JOHN
5	SUE

Two File

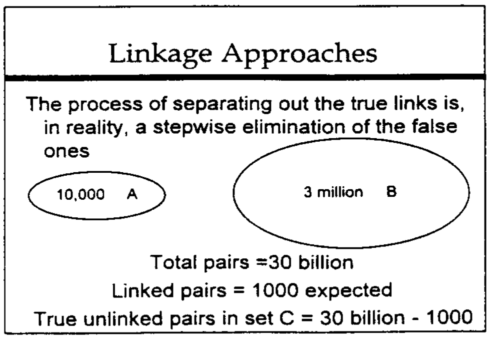
Table A		Table B
FRED	←——→	FRED
SUSAN	←——→	S
JOHN		



Theory of Record Linkage (1)

Set C has N_a x N_b record pairs (3x2=6)

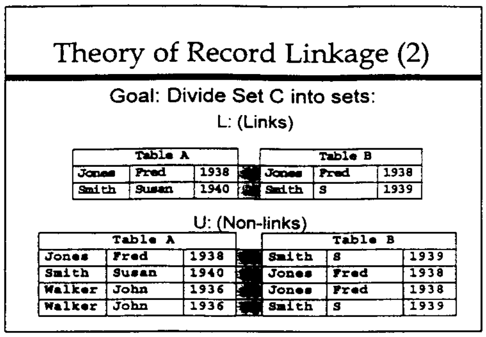
Table A				Table B		
Jones	Fred	1938		Jones	Fred	1938
Jones	Fred	1938		Smith	S	1939
Smith	Susan	1940		Jones	Fred	1938
Smith	Susan	1940		Smith	S	1939
Walker	John	1936		Jones	Fred	1938
Walker	John	1936		Smith	S	1939



Linkage Approaches

The process of separating out the true links is, in reality, a stepwise elimination of the false ones

Total pairs =30 billion
Linked pairs = 1000 expected
True unlinked pairs in set C = 30 billion - 1000



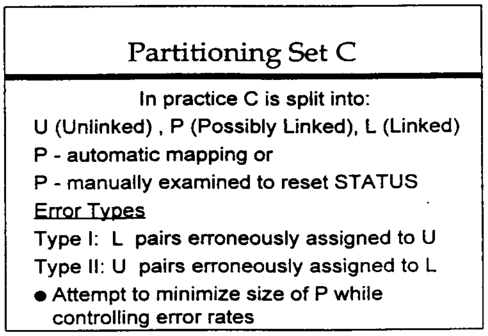
Theory of Record Linkage (2)

Goal: Divide Set C into sets:
L: (Links)

Table A				Table B		
Jones	Fred	1938		Jones	Fred	1938
Smith	Susan	1940		Smith	S	1939

U: (Non-links)

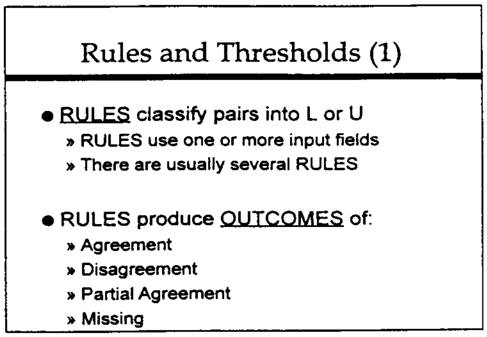
Table A				Table B		
Jones	Fred	1938		Smith	S	1939
Smith	Susan	1940		Jones	Fred	1938
Walker	John	1936		Jones	Fred	1938
Walker	John	1936		Smith	S	1939



Partitioning Set C

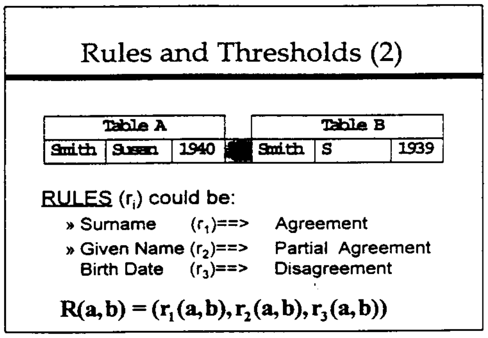
In practice C is split into:
U (Unlinked) , P (Possibly Linked), L (Linked)
P - automatic mapping or
P - manually examined to reset STATUS
Error Types
Type I: L pairs erroneously assigned to U
Type II: U pairs erroneously assigned to L

- Attempt to minimize size of P while controlling error rates



Rules and Thresholds (1)

- RULES classify pairs into L or U
  - » RULES use one or more input fields
  - » There are usually several RULES
- RULES produce OUTCOMES of:
  - » Agreement
  - » Disagreement
  - » Partial Agreement
  - » Missing



Rules and Thresholds (2)

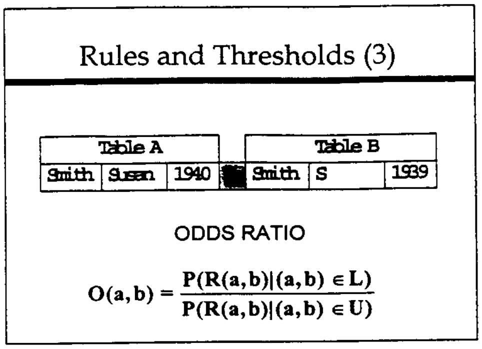
Table A				Table B		
Smith	Susan	1940		Smith	S	1939

RULES (r_i) could be:

» Surname (r_1)==> Agreement
» Given Name (r_2)==> Partial Agreement
Birth Date (r_3)==> Disagreement

$$R(a,b) = (r_1(a,b), r_2(a,b), r_3(a,b))$$

Slides Presentation (cont'd)

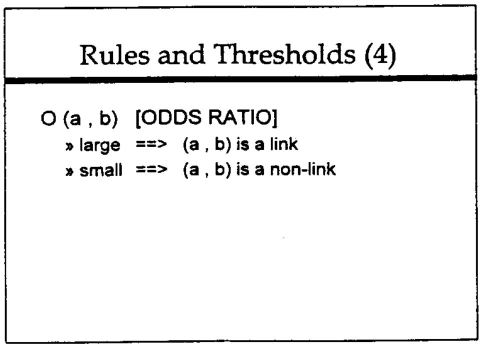


Rules and Thresholds (3)

Table A			Table B		
Smith	Susan	1940	Smith	S	1939

ODDS RATIO

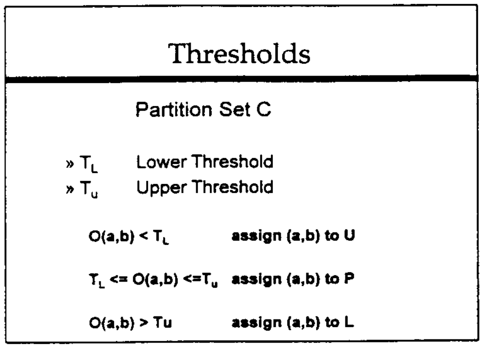
$$O(a,b) = \frac{P(R(a,b)|(a,b) \in L)}{P(R(a,b)|(a,b) \in U)}$$



Rules and Thresholds (4)

O (a , b) [ODDS RATIO]

- » large ==> (a , b) is a link
- » small ==> (a , b) is a non-link

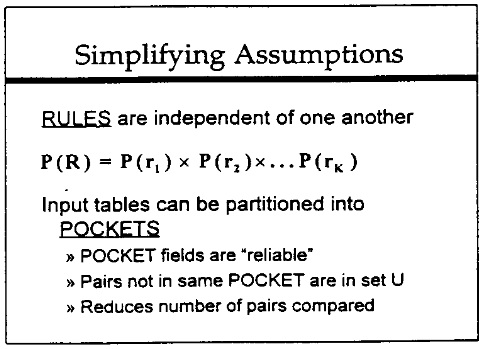


Thresholds

Partition Set C

- » T_L Lower Threshold
- » T_u Upper Threshold

$O(a,b) < T_L$	**assign (a,b) to U**
$T_L <= O(a,b) <= T_u$	**assign (a,b) to P**
O(a,b) > Tu	**assign (a,b) to L**



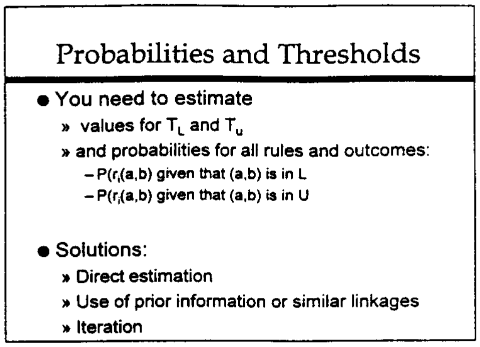
Simplifying Assumptions

RULES are independent of one another

$$P(R) = P(r_1) \times P(r_2) \times \ldots P(r_K)$$

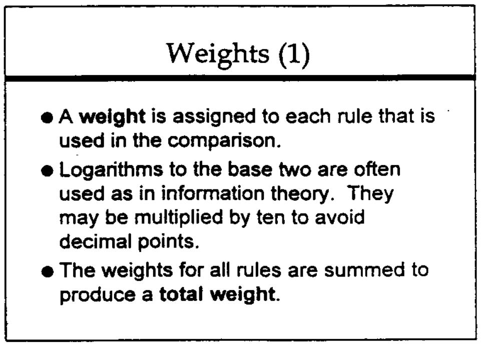
Input tables can be partitioned into POCKETS

- » POCKET fields are "reliable"
- » Pairs not in same POCKET are in set U
- » Reduces number of pairs compared



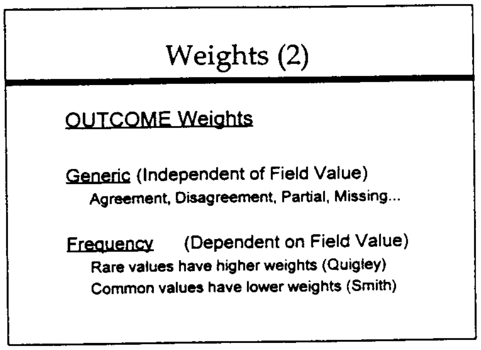
Probabilities and Thresholds

- You need to estimate
  - » values for T_L and T_u
  - » and probabilities for all rules and outcomes:
    - – $P(r_i(a,b)$ given that (a,b) is in L
    - – $P(r_i(a,b)$ given that (a,b) is in U
- Solutions:
  - » Direct estimation
  - » Use of prior information or similar linkages
  - » Iteration



Weights (1)

- A **weight** is assigned to each rule that is used in the comparison.
- Logarithms to the base two are often used as in information theory. They may be multiplied by ten to avoid decimal points.
- The weights for all rules are summed to produce a **total weight**.

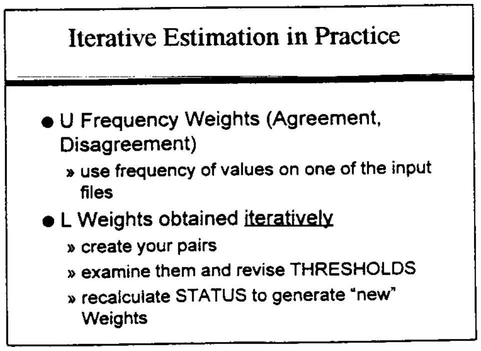


Weights (2)

OUTCOME Weights

Generic (Independent of Field Value)
Agreement, Disagreement, Partial, Missing...

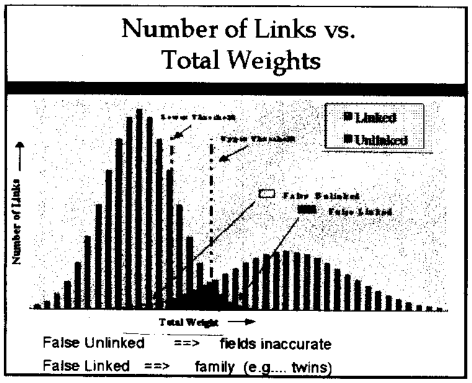
Frequency (Dependent on Field Value)
Rare values have higher weights (Quigley)
Common values have lower weights (Smith)



Iterative Estimation in Practice

- U Frequency Weights (Agreement, Disagreement)
  - » use frequency of values on one of the input files
- L Weights obtained iteratively
  - » create your pairs
  - » examine them and revise THRESHOLDS
  - » recalculate STATUS to generate "new" Weights

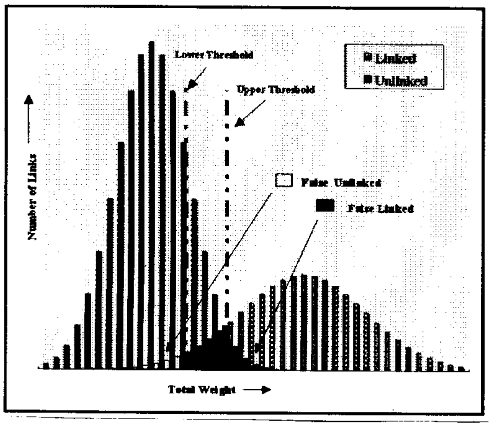
Slides Presentation (cont'd)

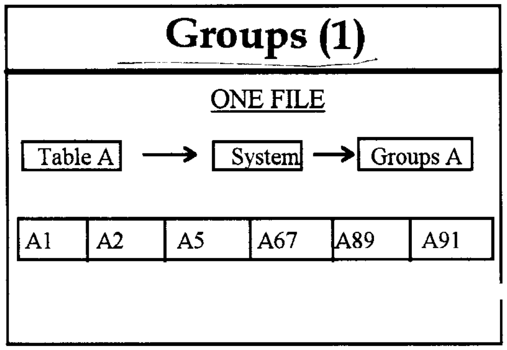


Number of Links vs. Total Weights

False Unlinked ==> fields inaccurate

False Linked ==> family (e.g.... twins)



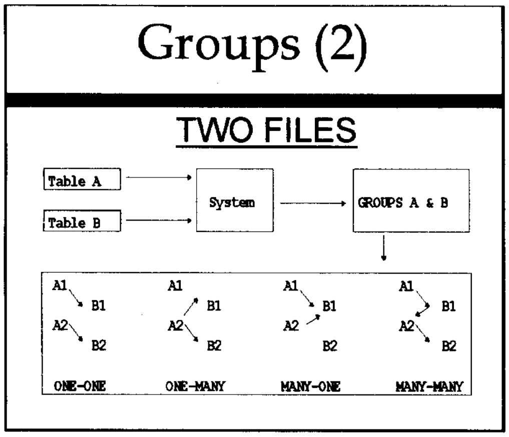


Groups (1)

ONE FILE

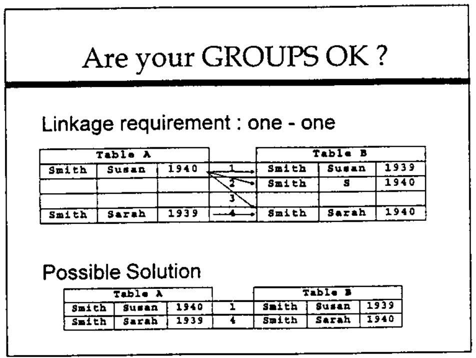
Table A → System → Groups A

A1	A2	A5	A67	A89	A91



Groups (2)

TWO FILES



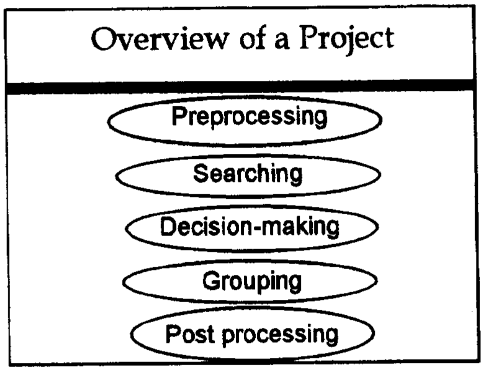
Are your GROUPS OK ?

Linkage requirement : one - one

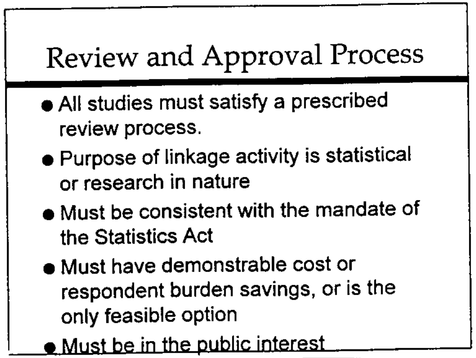
Table A				Table B		
Smith	Susan	1940	1	Smith	Susan	1939
			2	Smith	S	1940
			3			
Smith	Sarah	1939	4	Smith	Sarah	1940

Possible Solution

Table A				Table B		
Smith	Susan	1940	1	Smith	Susan	1939
Smith	Sarah	1939	4	Smith	Sarah	1940

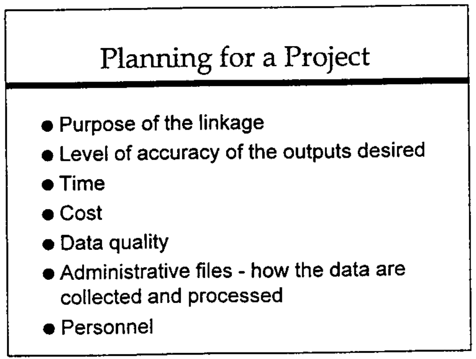


Overview of a Project



Review and Approval Process

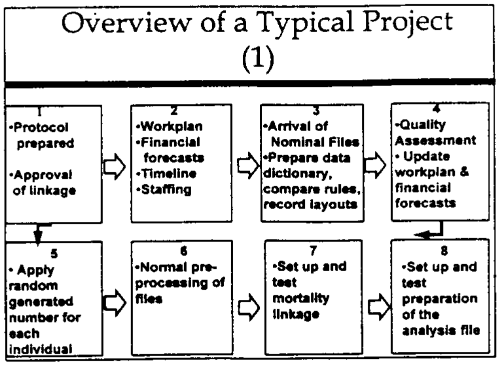
- All studies must satisfy a prescribed review process.
- Purpose of linkage activity is statistical or research in nature
- Must be consistent with the mandate of the Statistics Act
- Must have demonstrable cost or respondent burden savings, or is the only feasible option
- Must be in the public interest



Planning for a Project

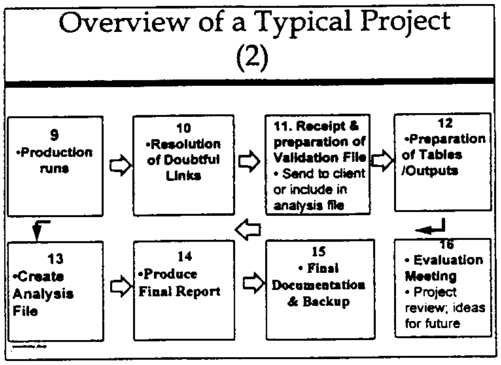
- Purpose of the linkage
- Level of accuracy of the outputs desired
- Time
- Cost
- Data quality
- Administrative files - how the data are collected and processed
- Personnel

Slides Presentation (cont'd)



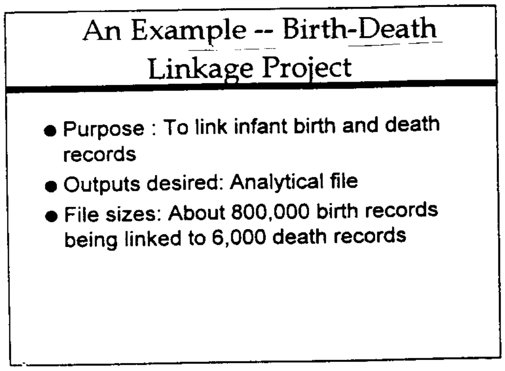
Overview of a Typical Project (1)

1
•Protocol prepared
•Approval of linkage
2
•Workplan
•Financial forecasts
•Timeline
•Staffing
3
•Arrival of Nominal Files
•Prepare data dictionary, compare rules, record layouts
4
•Quality Assessment
• Update workplan & financial forecasts
5
• Apply random generated number for each individual
6
•Normal pre-processing of files
7
•Set up and test mortality linkage
8
•Set up and test preparation of the analysis file



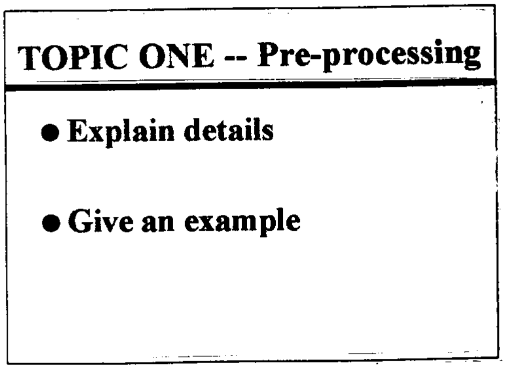
Overview of a Typical Project (2)

9
•Production runs
10
•Resolution of Doubtful Links
11. Receipt & preparation of Validation File
• Send to client or include in analysis file
12
•Preparation of Tables /Outputs
13
•Create Analysis File
14
•Produce Final Report
15
• Final Documentation & Backup
16
• Evaluation Meeting
• Project review; ideas for future



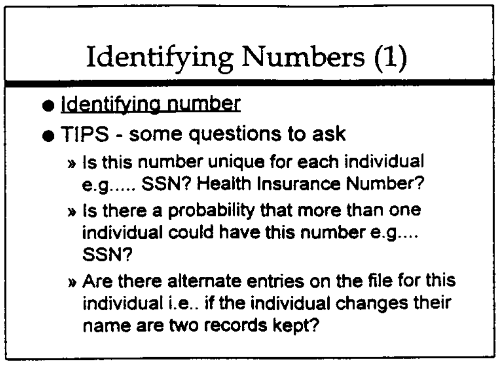
An Example -- Birth-Death Linkage Project

- Purpose : To link infant birth and death records
- Outputs desired: Analytical file
- File sizes: About 800,000 birth records being linked to 6,000 death records



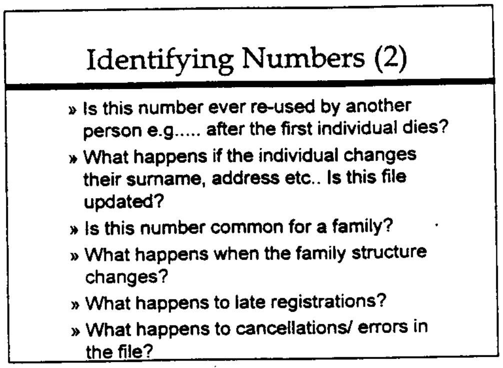
TOPIC ONE -- Pre-processing

- **Explain details**
- **Give an example**



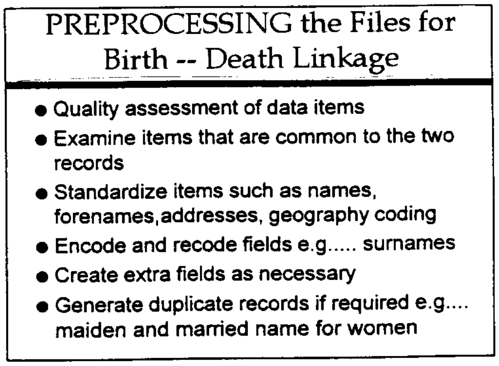
Identifying Numbers (1)

- Identifying number
- TIPS - some questions to ask
  - » Is this number unique for each individual e.g..... SSN? Health Insurance Number?
  - » Is there a probability that more than one individual could have this number e.g.... SSN?
  - » Are there alternate entries on the file for this individual i.e.. if the individual changes their name are two records kept?



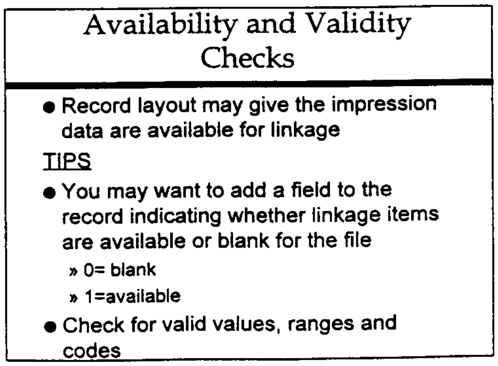
Identifying Numbers (2)

- » Is this number ever re-used by another person e.g..... after the first individual dies?
- » What happens if the individual changes their surname, address etc.. Is this file updated?
- » Is this number common for a family?
- » What happens when the family structure changes?
- » What happens to late registrations?
- » What happens to cancellations/ errors in the file?



PREPROCESSING the Files for Birth -- Death Linkage

- Quality assessment of data items
- Examine items that are common to the two records
- Standardize items such as names, forenames,addresses, geography coding
- Encode and recode fields e.g..... surnames
- Create extra fields as necessary
- Generate duplicate records if required e.g.... maiden and married name for women



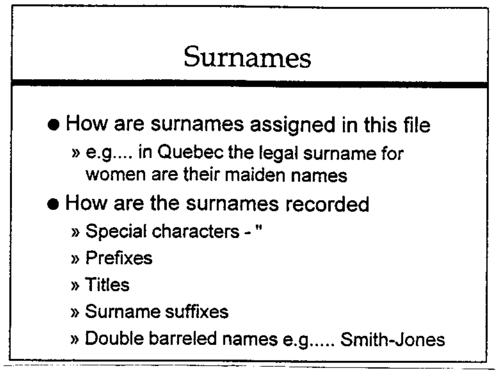
Availability and Validity Checks

- Record layout may give the impression data are available for linkage

TIPS

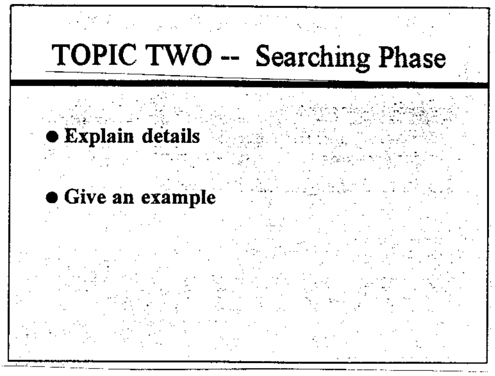
- You may want to add a field to the record indicating whether linkage items are available or blank for the file
  - » 0= blank
  - » 1=available
- Check for valid values, ranges and codes

Slides Presentation (cont'd)



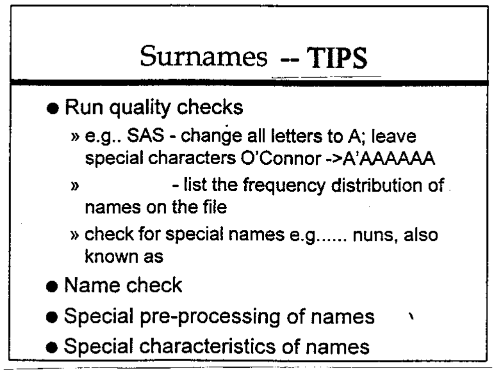
Surnames

- How are surnames assigned in this file
  - » e.g.... in Quebec the legal surname for women are their maiden names
- How are the surnames recorded
  - » Special characters - "
  - » Prefixes
  - » Titles
  - » Surname suffixes
  - » Double barreled names e.g..... Smith-Jones



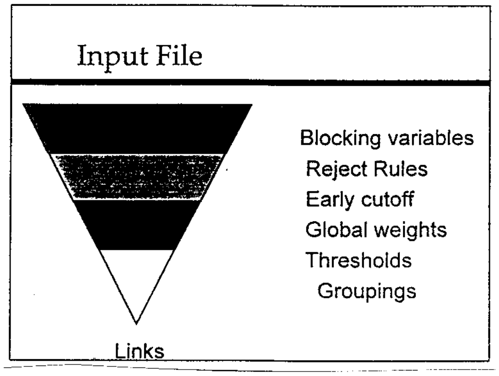
TOPIC TWO -- Searching Phase

- **Explain details**
- **Give an example**

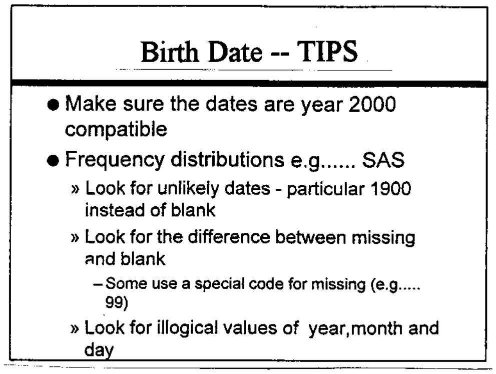


Surnames -- TIPS

- Run quality checks
  - » e.g.. SAS - change all letters to A; leave special characters O'Connor ->A'AAAAAA
  - » - list the frequency distribution of names on the file
  - » check for special names e.g...... nuns, also known as
- Name check
- Special pre-processing of names
- Special characteristics of names

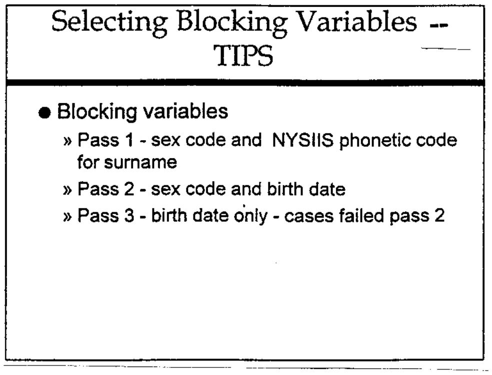


Input File



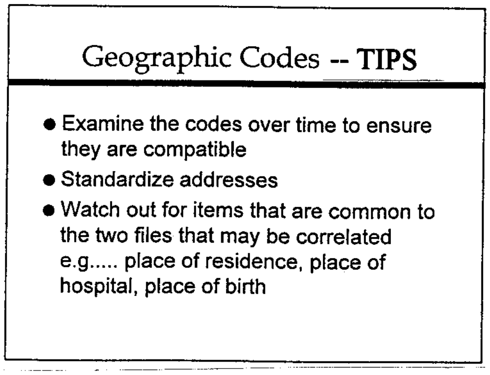
Birth Date -- TIPS

- Make sure the dates are year 2000 compatible
- Frequency distributions e.g...... SAS
  - » Look for unlikely dates - particular 1900 instead of blank
  - » Look for the difference between missing and blank
    - – Some use a special code for missing (e.g..... 99)
  - » Look for illogical values of year,month and day



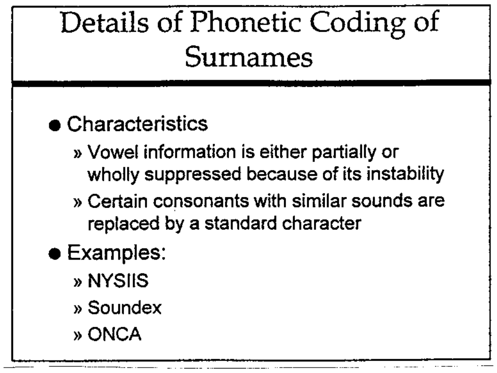
Selecting Blocking Variables -- TIPS

- Blocking variables
  - » Pass 1 - sex code and NYSIIS phonetic code for surname
  - » Pass 2 - sex code and birth date
  - » Pass 3 - birth date only - cases failed pass 2



Geographic Codes -- TIPS

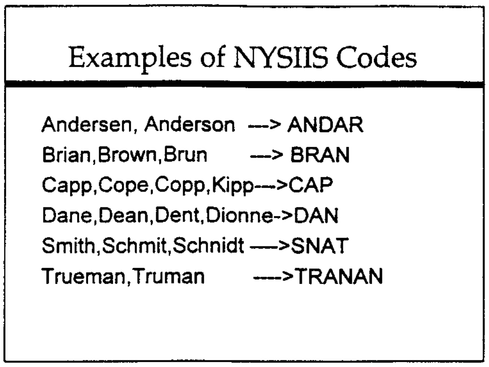
- Examine the codes over time to ensure they are compatible
- Standardize addresses
- Watch out for items that are common to the two files that may be correlated e.g..... place of residence, place of hospital, place of birth



Details of Phonetic Coding of Surnames

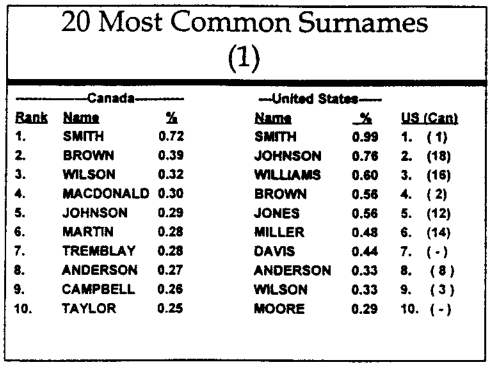
- Characteristics
  - » Vowel information is either partially or wholly suppressed because of its instability
  - » Certain consonants with similar sounds are replaced by a standard character
- Examples:
  - » NYSIIS
  - » Soundex
  - » ONCA

Slides Presentation (cont'd)



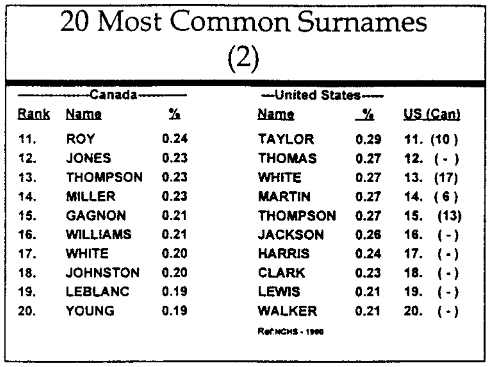
Examples of NYSIIS Codes

Andersen, Anderson —> ANDAR
Brian,Brown,Brun —> BRAN
Capp,Cope,Copp,Kipp—>CAP
Dane,Dean,Dent,Dionne->DAN
Smith,Schmit,Schnidt —>SNAT
Trueman,Truman —>TRANAN



20 Most Common Surnames (1)

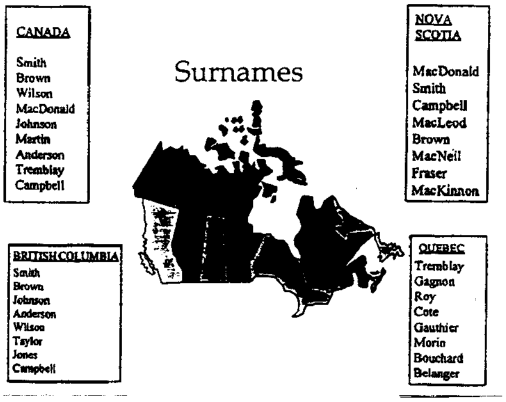
——Canada——			—United States—		
Rank	Name	%	Name	%	US (Can)
1.	SMITH	0.72	SMITH	0.99	1. (1)
2.	BROWN	0.39	JOHNSON	0.76	2. (18)
3.	WILSON	0.32	WILLIAMS	0.60	3. (16)
4.	MACDONALD	0.30	BROWN	0.56	4. (2)
5.	JOHNSON	0.29	JONES	0.56	5. (12)
6.	MARTIN	0.28	MILLER	0.48	6. (14)
7.	TREMBLAY	0.28	DAVIS	0.44	7. (-)
8.	ANDERSON	0.27	ANDERSON	0.33	8. (8)
9.	CAMPBELL	0.26	WILSON	0.33	9. (3)
10.	TAYLOR	0.25	MOORE	0.29	10. (-)



20 Most Common Surnames (2)

——Canada——			—United States—		
Rank	Name	%	Name	%	US (Can)
11.	ROY	0.24	TAYLOR	0.29	11. (10)
12.	JONES	0.23	THOMAS	0.27	12. (-)
13.	THOMPSON	0.23	WHITE	0.27	13. (17)
14.	MILLER	0.23	MARTIN	0.27	14. (6)
15.	GAGNON	0.21	THOMPSON	0.27	15. (13)
16.	WILLIAMS	0.21	JACKSON	0.26	16. (-)
17.	WHITE	0.20	HARRIS	0.24	17. (-)
18.	JOHNSTON	0.20	CLARK	0.23	18. (-)
19.	LEBLANC	0.19	LEWIS	0.21	19. (-)
20.	YOUNG	0.19	WALKER	0.21	20. (-)

Ref:NCHS - 1990



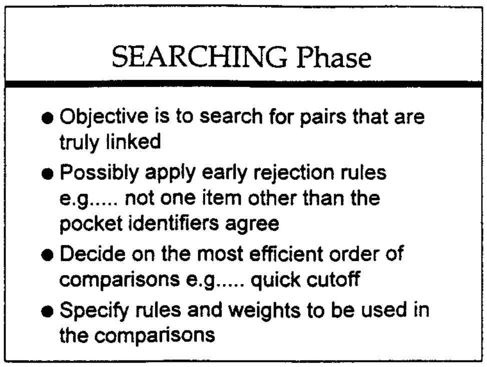
Surnames

CANADA
Smith
Brown
Wilson
MacDonald
Johnson
Martin
Anderson
Tremblay
Campbell

BRITISH COLUMBIA
Smith
Brown
Johnson
Anderson
Wilson
Taylor
Jones
Campbell

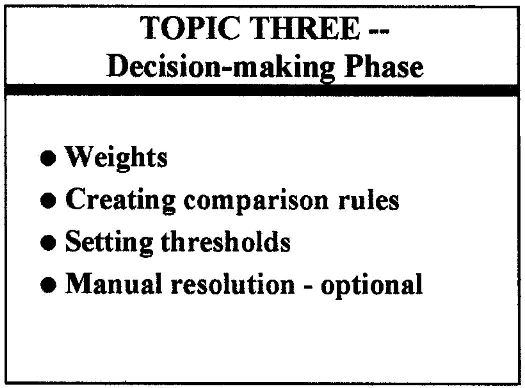
NOVA SCOTIA
MacDonald
Smith
Campbell
MacLeod
Brown
MacNeil
Fraser
MacKinnon

QUEBEC
Tremblay
Gagnon
Roy
Cote
Gauthier
Morin
Bouchard
Belanger



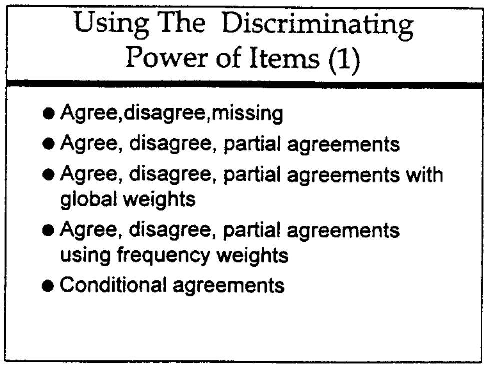
SEARCHING Phase

- Objective is to search for pairs that are truly linked
- Possibly apply early rejection rules e.g..... not one item other than the pocket identifiers agree
- Decide on the most efficient order of comparisons e.g..... quick cutoff
- Specify rules and weights to be used in the comparisons



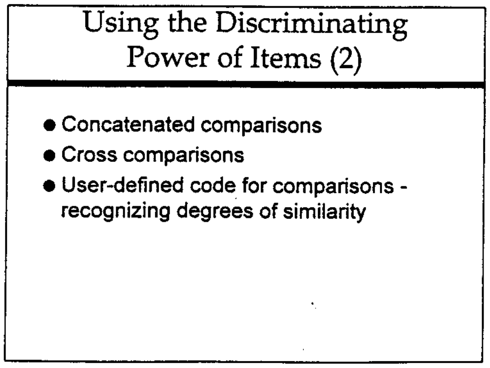
TOPIC THREE -- Decision-making Phase

- **Weights**
- **Creating comparison rules**
- **Setting thresholds**
- **Manual resolution - optional**



Using The Discriminating Power of Items (1)

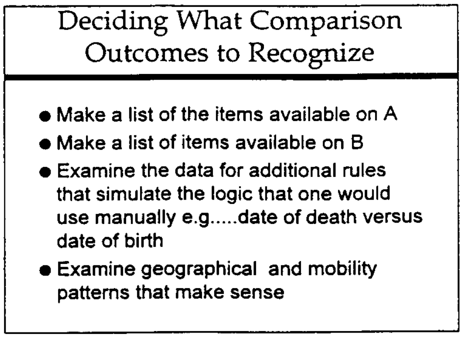
- Agree,disagree,missing
- Agree, disagree, partial agreements
- Agree, disagree, partial agreements with global weights
- Agree, disagree, partial agreements using frequency weights
- Conditional agreements



Using the Discriminating Power of Items (2)

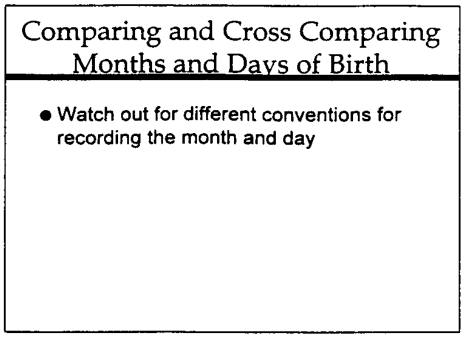
- Concatenated comparisons
- Cross comparisons
- User-defined code for comparisons - recognizing degrees of similarity

Slides Presentation (cont'd)



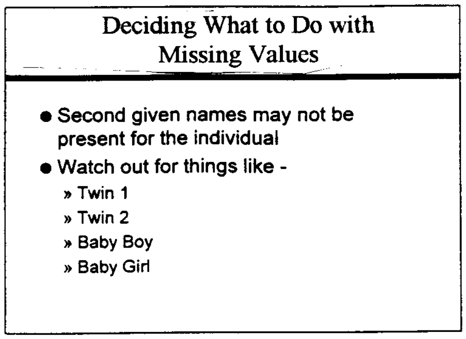
Deciding What Comparison Outcomes to Recognize

- Make a list of the items available on A
- Make a list of items available on B
- Examine the data for additional rules that simulate the logic that one would use manually e.g.....date of death versus date of birth
- Examine geographical and mobility patterns that make sense



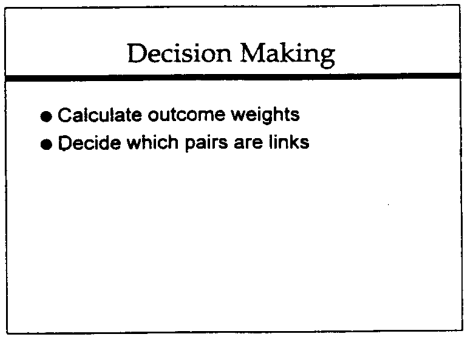
Comparing and Cross Comparing Months and Days of Birth

- Watch out for different conventions for recording the month and day



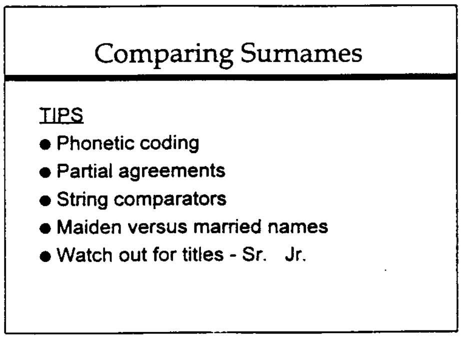
Deciding What to Do with Missing Values

- Second given names may not be present for the individual
- Watch out for things like -
  - » Twin 1
  - » Twin 2
  - » Baby Boy
  - » Baby Girl



Decision Making

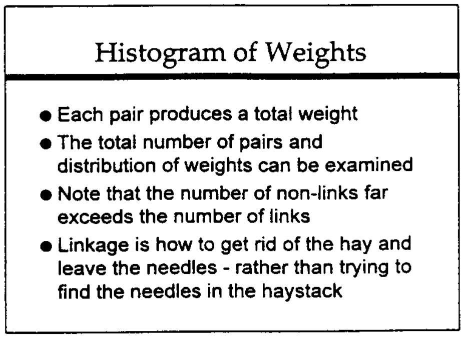
- Calculate outcome weights
- Decide which pairs are links



Comparing Surnames

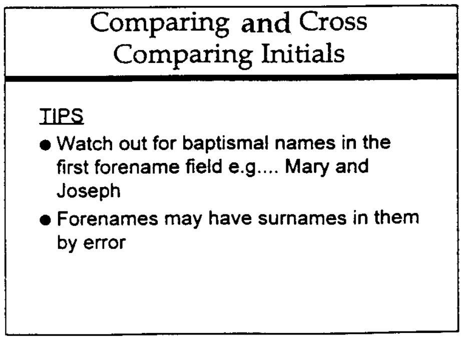
TIPS

- Phonetic coding
- Partial agreements
- String comparators
- Maiden versus married names
- Watch out for titles - Sr. Jr.



Histogram of Weights

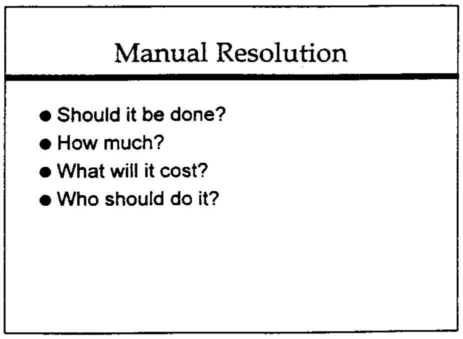
- Each pair produces a total weight
- The total number of pairs and distribution of weights can be examined
- Note that the number of non-links far exceeds the number of links
- Linkage is how to get rid of the hay and leave the needles - rather than trying to find the needles in the haystack



Comparing and Cross Comparing Initials

TIPS

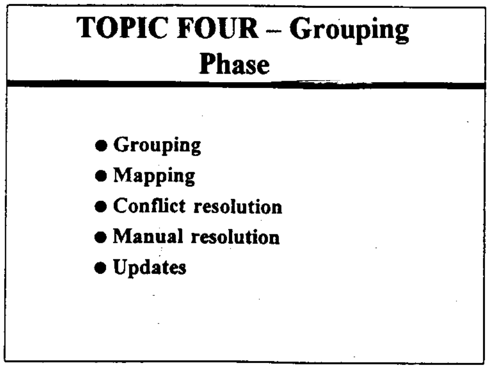
- Watch out for baptismal names in the first forename field e.g.... Mary and Joseph
- Forenames may have surnames in them by error



Manual Resolution

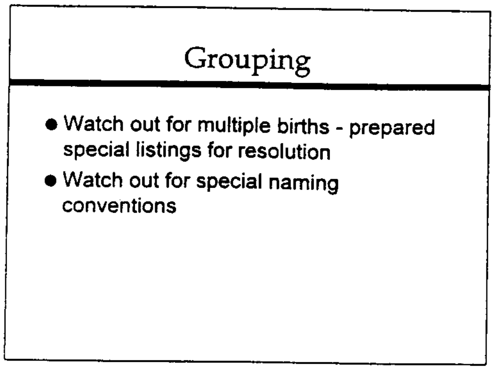
- Should it be done?
- How much?
- What will it cost?
- Who should do it?

Slides Presentation (cont'd)



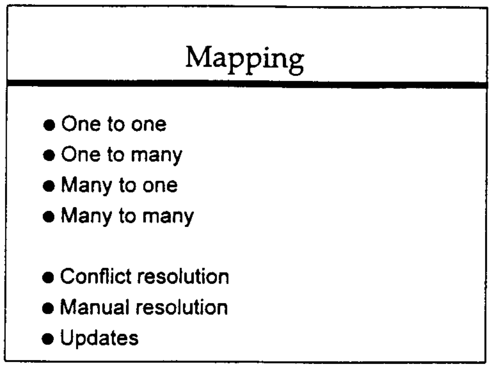
TOPIC FOUR – Grouping Phase

- **Grouping**
- **Mapping**
- **Conflict resolution**
- **Manual resolution**
- **Updates**



Grouping

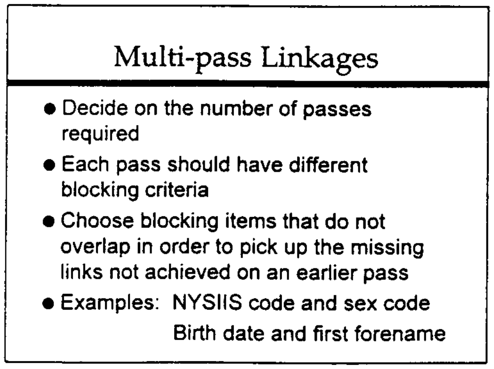
- Watch out for multiple births - prepared special listings for resolution
- Watch out for special naming conventions



Mapping

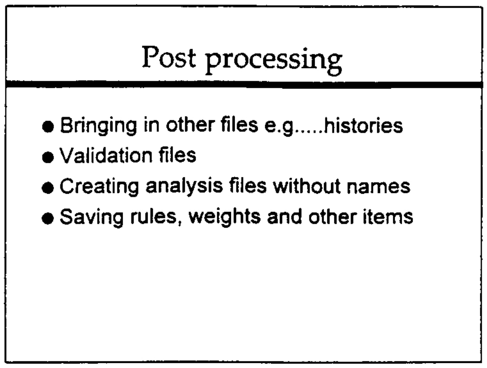
- One to one
- One to many
- Many to one
- Many to many

- Conflict resolution
- Manual resolution
- Updates



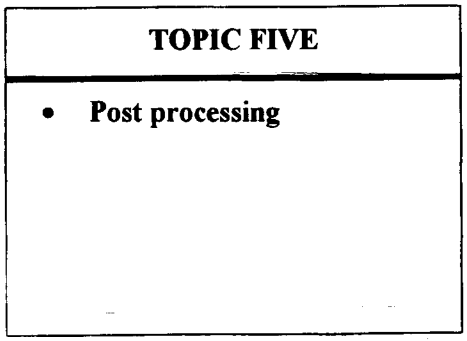
Multi-pass Linkages

- Decide on the number of passes required
- Each pass should have different blocking criteria
- Choose blocking items that do not overlap in order to pick up the missing links not achieved on an earlier pass
- Examples: NYSIIS code and sex code
  Birth date and first forename



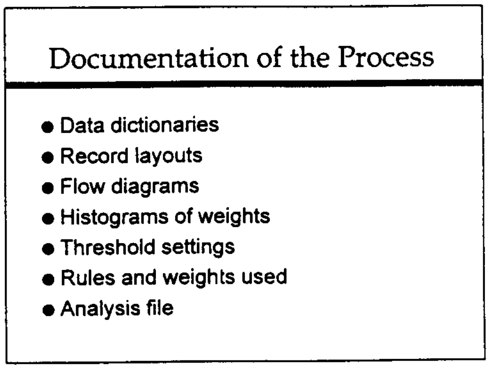
Post processing

- Bringing in other files e.g.....histories
- Validation files
- Creating analysis files without names
- Saving rules, weights and other items



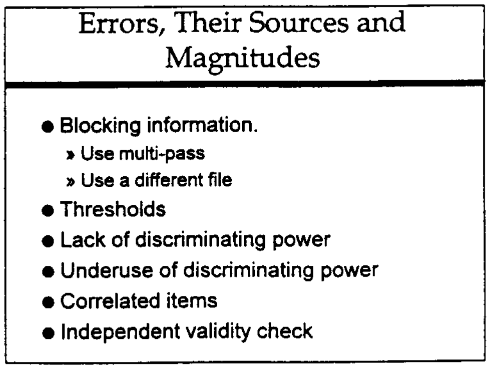
TOPIC FIVE

- **Post processing**



Documentation of the Process

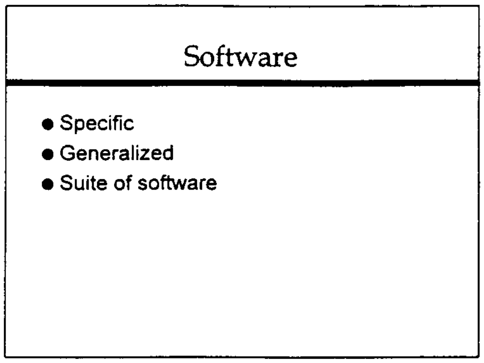
- Data dictionaries
- Record layouts
- Flow diagrams
- Histograms of weights
- Threshold settings
- Rules and weights used
- Analysis file



Errors, Their Sources and Magnitudes

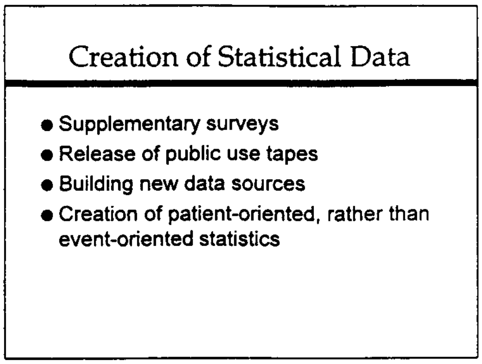
- Blocking information.
  - » Use multi-pass
  - » Use a different file
- Thresholds
- Lack of discriminating power
- Underuse of discriminating power
- Correlated items
- Independent validity check

Slides Presentation (cont'd)



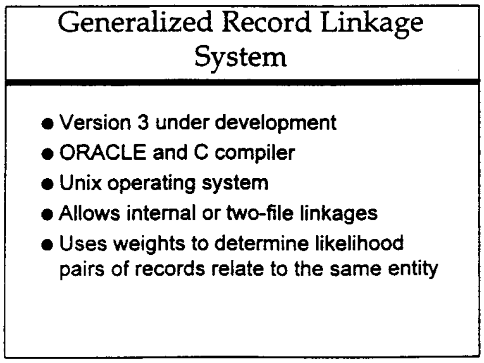
Software

- Specific
- Generalized
- Suite of software



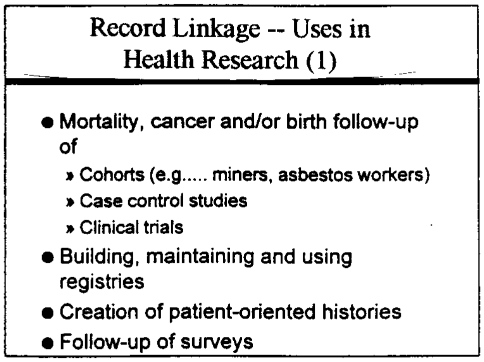
Creation of Statistical Data

- Supplementary surveys
- Release of public use tapes
- Building new data sources
- Creation of patient-oriented, rather than event-oriented statistics



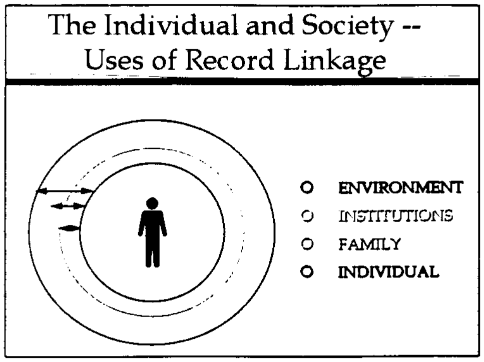
Generalized Record Linkage System

- Version 3 under development
- ORACLE and C compiler
- Unix operating system
- Allows internal or two-file linkages
- Uses weights to determine likelihood pairs of records relate to the same entity

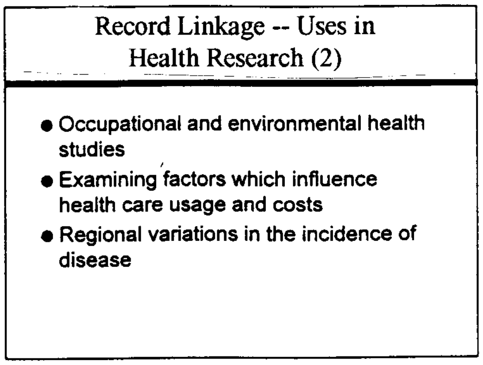


Record Linkage -- Uses in Health Research (1)

- Mortality, cancer and/or birth follow-up of
  - Cohorts (e.g..... miners, asbestos workers)
  - Case control studies
  - Clinical trials
- Building, maintaining and using registries
- Creation of patient-oriented histories
- Follow-up of surveys

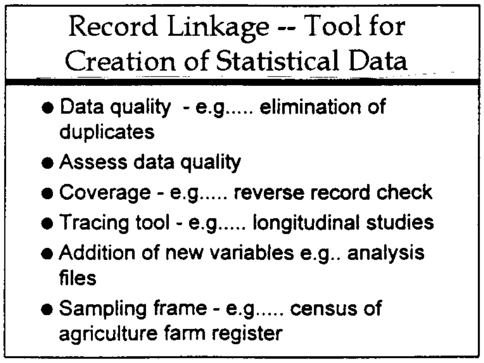


The Individual and Society -- Uses of Record Linkage



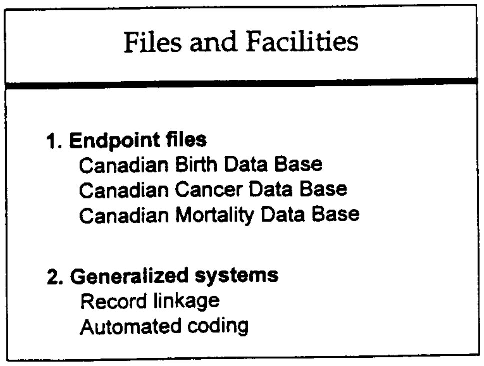
Record Linkage -- Uses in Health Research (2)

- Occupational and environmental health studies
- Examining factors which influence health care usage and costs
- Regional variations in the incidence of disease



Record Linkage -- Tool for Creation of Statistical Data

- Data quality - e.g..... elimination of duplicates
- Assess data quality
- Coverage - e.g..... reverse record check
- Tracing tool - e.g..... longitudinal studies
- Addition of new variables e.g.. analysis files
- Sampling frame - e.g..... census of agriculture farm register



Files and Facilities

1. **Endpoint files**
   Canadian Birth Data Base
   Canadian Cancer Data Base
   Canadian Mortality Data Base

2. **Generalized systems**
   Record linkage
   Automated coding

Slides Presentation (cont'd)

Use of Record Linkage in Cancer Registries

- Creation of cancer registries
- Maintaining cancer registries
- Death clearance of cancer registries
- Evaluating the quality of registries
- Ascertainment of new death certificate only cases
- Replacing or partially replacing active follow-up of patients
- Carrying out cohort studies
- Follow-up of clinical trials and screening programs

Follow-up of National Breast Screening Program Cohort

OBJECTIVES

To follow-up women to determine the dates and causes of death

To confirm the diagnosis and cancer incidence of the study population

Number of Individuals:
90,000 females

YEARS

Cancer years:
1977-1993

Death years:
1980-1988
1989-1993

ORGANIZATIONS
University of Toronto
Statistics Canada

Advantages of Record Linkage in Cancer Registries

- Reduces respondent burden
- Improves accuracy
- Reduced follow-up costs
- Refines detection and measurement of mortality and cancer rates for particular cohorts

Death Clearance of the Nova Scotia Cancer Registry

OBJECTIVES

- **To calculate survival rates for persons with cancer in Nova Scotia**
- **To add "death certificate only" cases to the Nova Scotia Cancer Registry**

Number of Records:
About 79,000 cancer records relating to about 60,000 individuals were linked to 3.6 million individual deaths

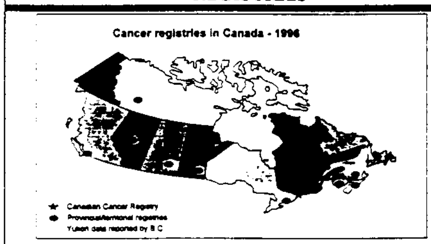
YEARS

Cancer years:
1969-1988

Death years:
1969-1989

ORGANIZATIONS
- **Health Canada**
- **Nova Scotia Cancer Registry**
- **Statistics Canada**

USE OF RECORD LINKAGE IN BUILDING, MAINTAINING AND USING CANCER REGISTRIES


Cancer registries in Canada - 1996
Canadian Cancer Registry
Provincial/territorial registries
Yukon data reported by B C

Canada

Occupational Studies

- Ontario cancer study
  - » Feasibility study in seven Ontario regions
  - » Linkage to Ontario cancer registry

Follow-up of Childhood Cancers

OBJECTIVES

- To examine the problems and risks facing Canadian children with respect to cancer

Number of Individuals:
Approximately 17,000
(including 1985-1991 Ontario cases)

YEARS

Cancer years:
1969-1988

Death years:
1969-1991

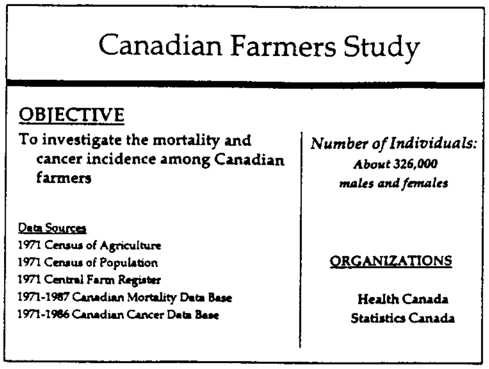
ORGANIZATIONS

Health Canada
Statistics Canada
Provincial cancer agencies

Key Elements in a Typical Study

1. Exposed population	**5. Levels of exposure**
2. Comparison group	**6. Time course of response**
3. Hazard identification	**7. Confounding factors**
4. Endpoints	

Slides Presentation (cont'd)



Canadian Farmers Study

OBJECTIVE

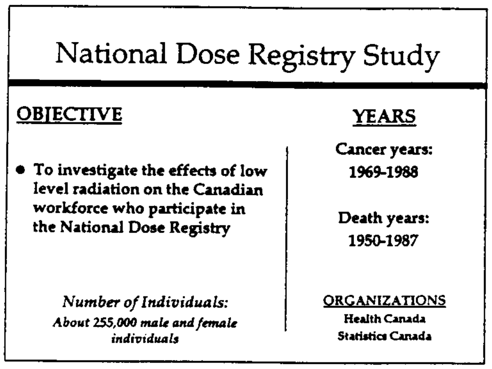
To investigate the mortality and cancer incidence among Canadian farmers

Data Sources
1971 Census of Agriculture
1971 Census of Population
1971 Central Farm Register
1971-1987 Canadian Mortality Data Base
1971-1986 Canadian Cancer Data Base

Number of Individuals:
About 326,000 males and females

ORGANIZATIONS

Health Canada
Statistics Canada



National Dose Registry Study

OBJECTIVE

- To investigate the effects of low level radiation on the Canadian workforce who participate in the National Dose Registry

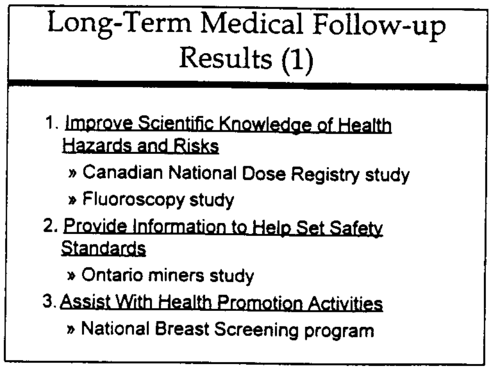
Number of Individuals:
About 255,000 male and female individuals

YEARS

Cancer years:
1969-1988

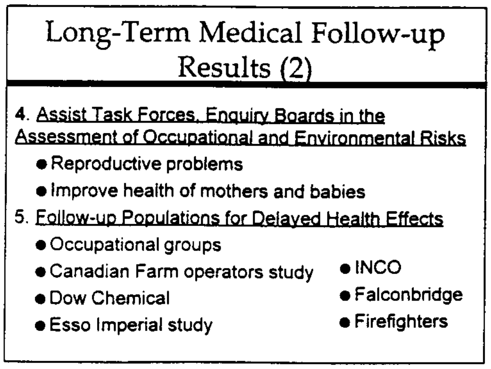
Death years:
1950-1987

ORGANIZATIONS
Health Canada
Statistics Canada



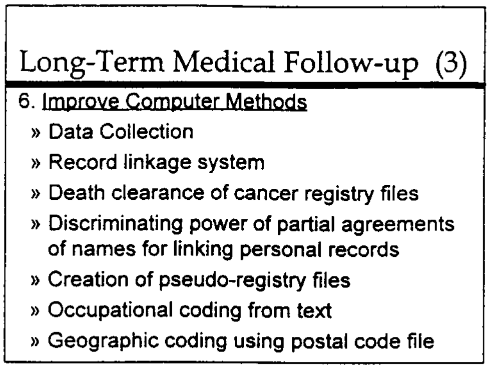
Long-Term Medical Follow-up Results (1)

1. Improve Scientific Knowledge of Health Hazards and Risks
   - » Canadian National Dose Registry study
   - » Fluoroscopy study
2. Provide Information to Help Set Safety Standards
   - » Ontario miners study
3. Assist With Health Promotion Activities
   - » National Breast Screening program



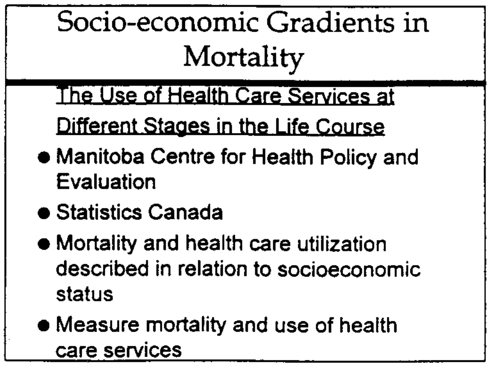
Long-Term Medical Follow-up Results (2)

4. Assist Task Forces, Enquiry Boards in the Assessment of Occupational and Environmental Risks
   - ● Reproductive problems
   - ● Improve health of mothers and babies
5. Follow-up Populations for Delayed Health Effects
   - ● Occupational groups
   - ● Canadian Farm operators study
   - ● Dow Chemical
   - ● Esso Imperial study
   - ● INCO
   - ● Falconbridge
   - ● Firefighters



Long-Term Medical Follow-up (3)

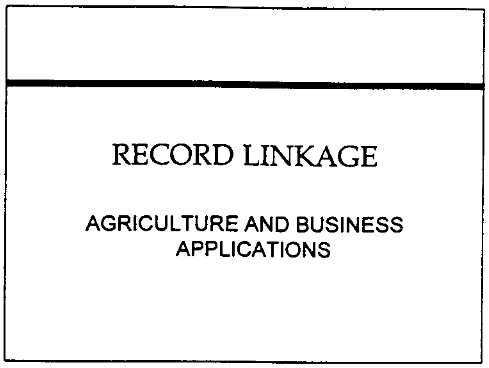
6. Improve Computer Methods
   - » Data Collection
   - » Record linkage system
   - » Death clearance of cancer registry files
   - » Discriminating power of partial agreements of names for linking personal records
   - » Creation of pseudo-registry files
   - » Occupational coding from text
   - » Geographic coding using postal code file



Socio-economic Gradients in Mortality

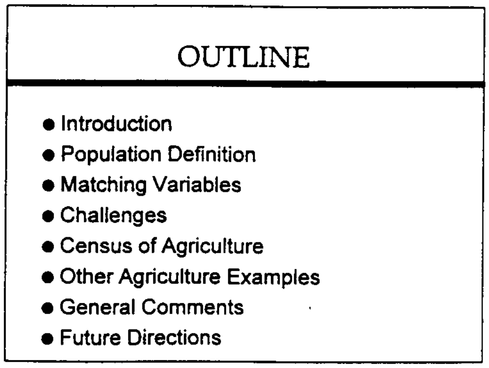
The Use of Health Care Services at Different Stages in the Life Course

- ● Manitoba Centre for Health Policy and Evaluation
- ● Statistics Canada
- ● Mortality and health care utilization described in relation to socioeconomic status
- ● Measure mortality and use of health care services



RECORD LINKAGE

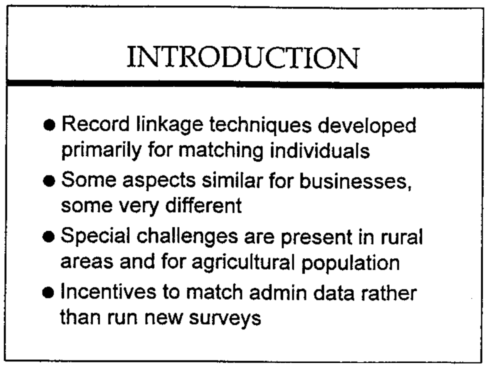
AGRICULTURE AND BUSINESS APPLICATIONS



OUTLINE

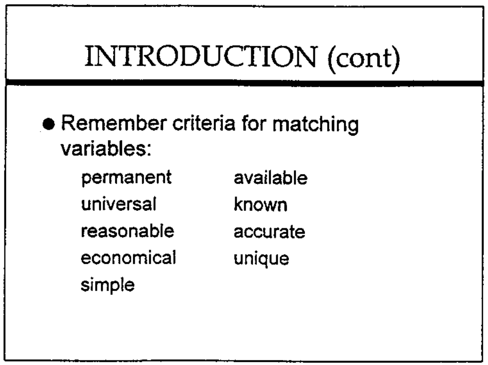
- ● Introduction
- ● Population Definition
- ● Matching Variables
- ● Challenges
- ● Census of Agriculture
- ● Other Agriculture Examples
- ● General Comments
- ● Future Directions

Slides Presentation (cont'd)



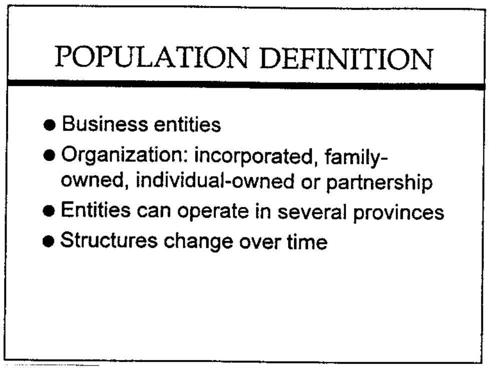
INTRODUCTION

- Record linkage techniques developed primarily for matching individuals
- Some aspects similar for businesses, some very different
- Special challenges are present in rural areas and for agricultural population
- Incentives to match admin data rather than run new surveys



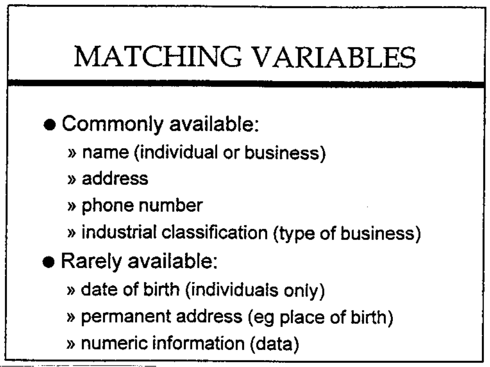
INTRODUCTION (cont)

- Remember criteria for matching variables:
  - permanent
  - universal
  - reasonable
  - economical
  - simple
  - available
  - known
  - accurate
  - unique



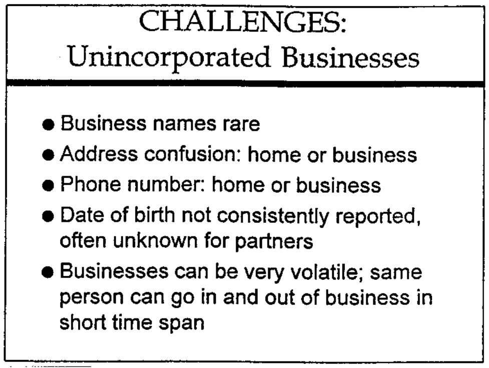
POPULATION DEFINITION

- Business entities
- Organization: incorporated, family-owned, individual-owned or partnership
- Entities can operate in several provinces
- Structures change over time



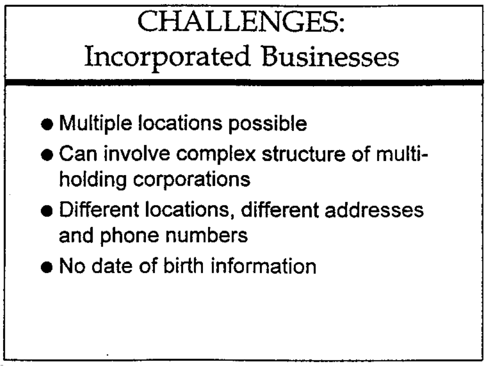
MATCHING VARIABLES

- Commonly available:
  - » name (individual or business)
  - » address
  - » phone number
  - » industrial classification (type of business)
- Rarely available:
  - » date of birth (individuals only)
  - » permanent address (eg place of birth)
  - » numeric information (data)



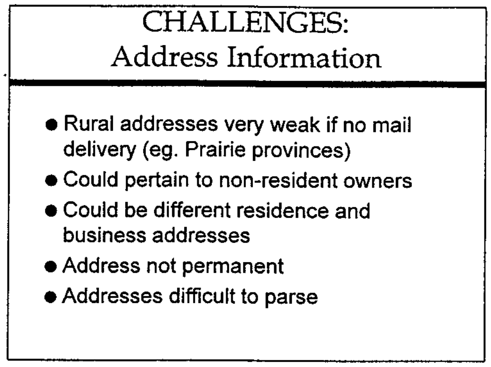
CHALLENGES: Unincorporated Businesses

- Business names rare
- Address confusion: home or business
- Phone number: home or business
- Date of birth not consistently reported, often unknown for partners
- Businesses can be very volatile; same person can go in and out of business in short time span



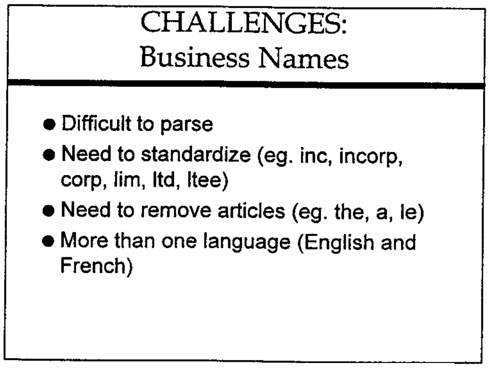
CHALLENGES: Incorporated Businesses

- Multiple locations possible
- Can involve complex structure of multi-holding corporations
- Different locations, different addresses and phone numbers
- No date of birth information



CHALLENGES: Address Information

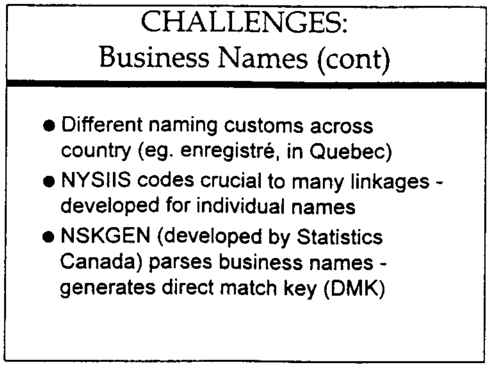
- Rural addresses very weak if no mail delivery (eg. Prairie provinces)
- Could pertain to non-resident owners
- Could be different residence and business addresses
- Address not permanent
- Addresses difficult to parse



CHALLENGES: Business Names

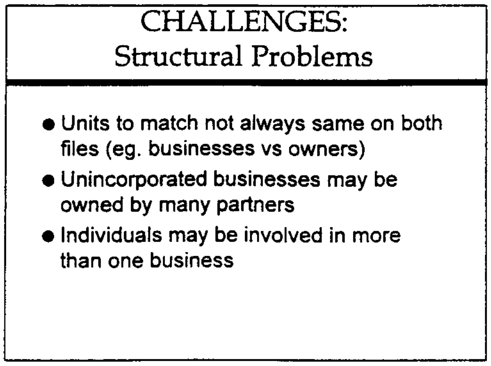
- Difficult to parse
- Need to standardize (eg. inc, incorp, corp, lim, ltd, ltee)
- Need to remove articles (eg. the, a, le)
- More than one language (English and French)

Slides Presentation (cont'd)



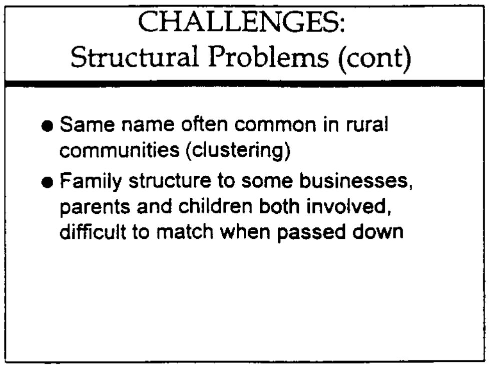
CHALLENGES: Business Names (cont)

- Different naming customs across country (eg. enregistré, in Quebec)
- NYSIIS codes crucial to many linkages - developed for individual names
- NSKGEN (developed by Statistics Canada) parses business names - generates direct match key (DMK)



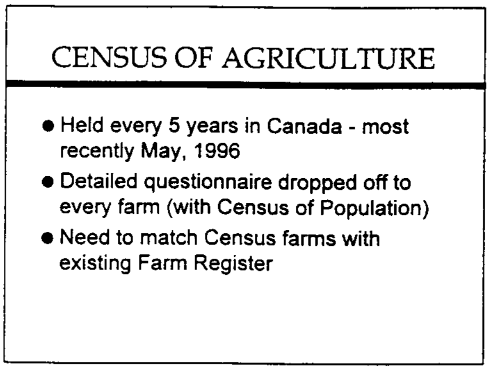
CHALLENGES: Structural Problems

- Units to match not always same on both files (eg. businesses vs owners)
- Unincorporated businesses may be owned by many partners
- Individuals may be involved in more than one business



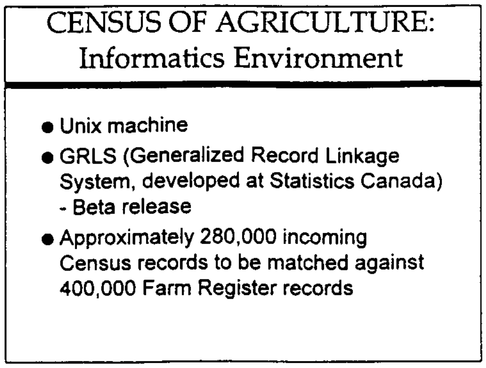
CHALLENGES: Structural Problems (cont)

- Same name often common in rural communities (clustering)
- Family structure to some businesses, parents and children both involved, difficult to match when passed down



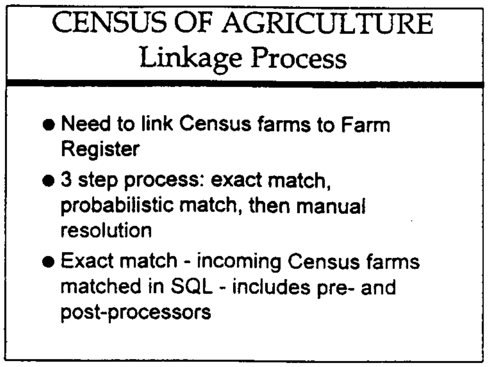
CENSUS OF AGRICULTURE

- Held every 5 years in Canada - most recently May, 1996
- Detailed questionnaire dropped off to every farm (with Census of Population)
- Need to match Census farms with existing Farm Register



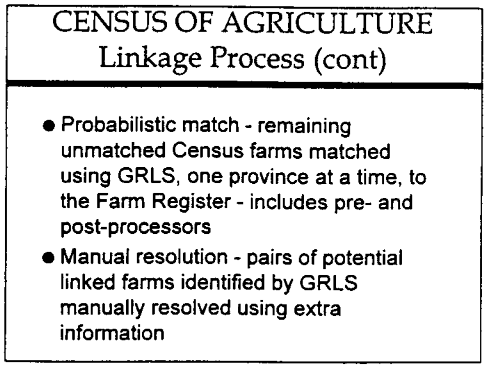
CENSUS OF AGRICULTURE: Informatics Environment

- Unix machine
- GRLS (Generalized Record Linkage System, developed at Statistics Canada) - Beta release
- Approximately 280,000 incoming Census records to be matched against 400,000 Farm Register records



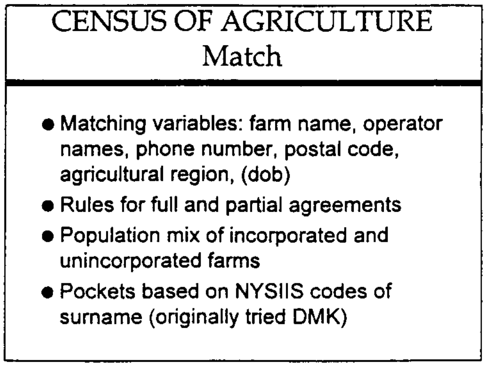
CENSUS OF AGRICULTURE Linkage Process

- Need to link Census farms to Farm Register
- 3 step process: exact match, probabilistic match, then manual resolution
- Exact match - incoming Census farms matched in SQL - includes pre- and post-processors



CENSUS OF AGRICULTURE Linkage Process (cont)

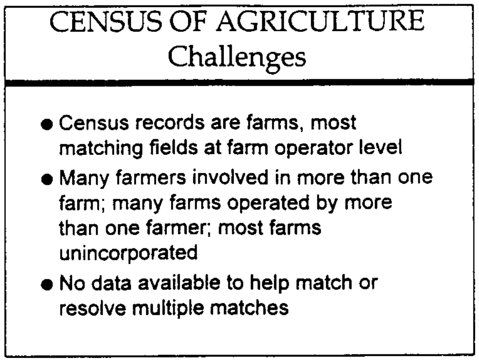
- Probabilistic match - remaining unmatched Census farms matched using GRLS, one province at a time, to the Farm Register - includes pre- and post-processors
- Manual resolution - pairs of potential linked farms identified by GRLS manually resolved using extra information



CENSUS OF AGRICULTURE Match

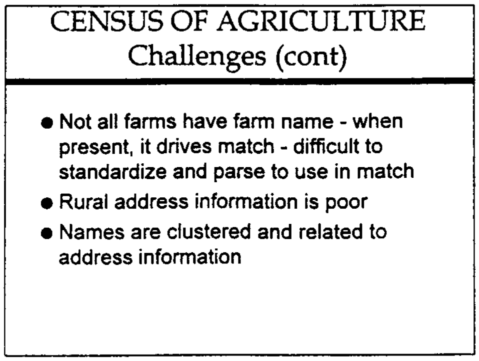
- Matching variables: farm name, operator names, phone number, postal code, agricultural region, (dob)
- Rules for full and partial agreements
- Population mix of incorporated and unincorporated farms
- Pockets based on NYSIIS codes of surname (originally tried DMK)

Slides Presentation (cont'd)



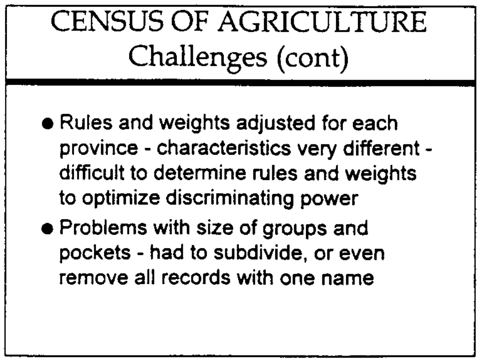
CENSUS OF AGRICULTURE Challenges

- Census records are farms, most matching fields at farm operator level
- Many farmers involved in more than one farm; many farms operated by more than one farmer; most farms unincorporated
- No data available to help match or resolve multiple matches



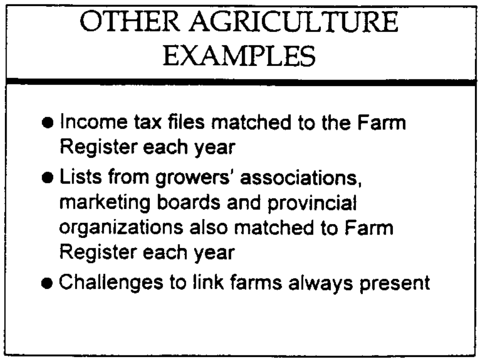
CENSUS OF AGRICULTURE Challenges (cont)

- Not all farms have farm name - when present, it drives match - difficult to standardize and parse to use in match
- Rural address information is poor
- Names are clustered and related to address information



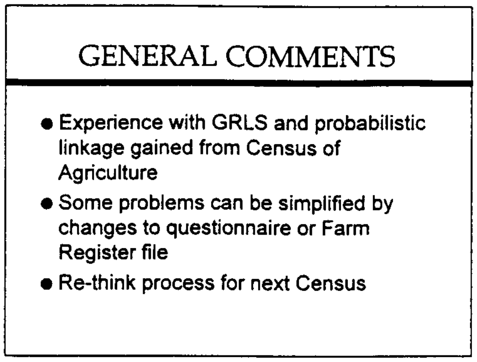
CENSUS OF AGRICULTURE Challenges (cont)

- Rules and weights adjusted for each province - characteristics very different - difficult to determine rules and weights to optimize discriminating power
- Problems with size of groups and pockets - had to subdivide, or even remove all records with one name



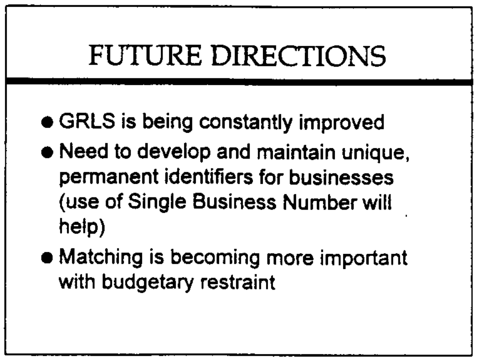
OTHER AGRICULTURE EXAMPLES

- Income tax files matched to the Farm Register each year
- Lists from growers' associations, marketing boards and provincial organizations also matched to Farm Register each year
- Challenges to link farms always present



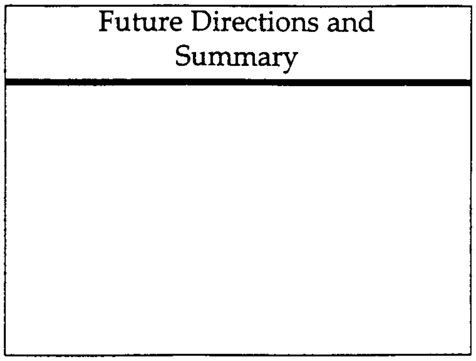
GENERAL COMMENTS

- Experience with GRLS and probabilistic linkage gained from Census of Agriculture
- Some problems can be simplified by changes to questionnaire or Farm Register file
- Re-think process for next Census

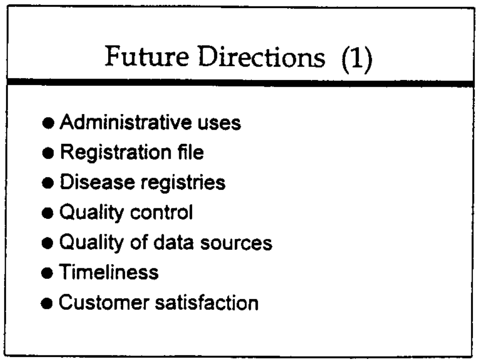


FUTURE DIRECTIONS

- GRLS is being constantly improved
- Need to develop and maintain unique, permanent identifiers for businesses (use of Single Business Number will help)
- Matching is becoming more important with budgetary restraint



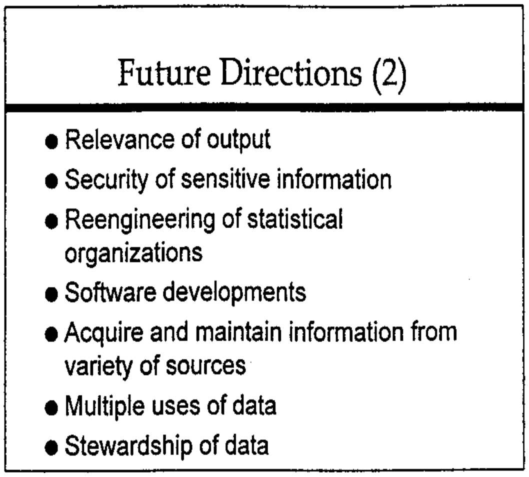
Future Directions and Summary



Future Directions (1)

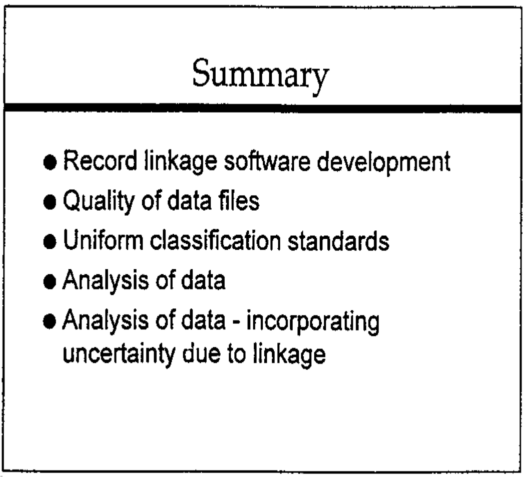
- Administrative uses
- Registration file
- Disease registries
- Quality control
- Quality of data sources
- Timeliness
- Customer satisfaction

Slides Presentation (cont'd)



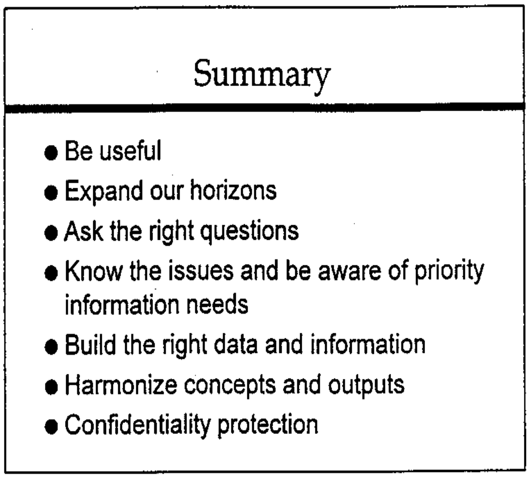
Future Directions (2)

- Relevance of output
- Security of sensitive information
- Reengineering of statistical organizations
- Software developments
- Acquire and maintain information from variety of sources
- Multiple uses of data
- Stewardship of data



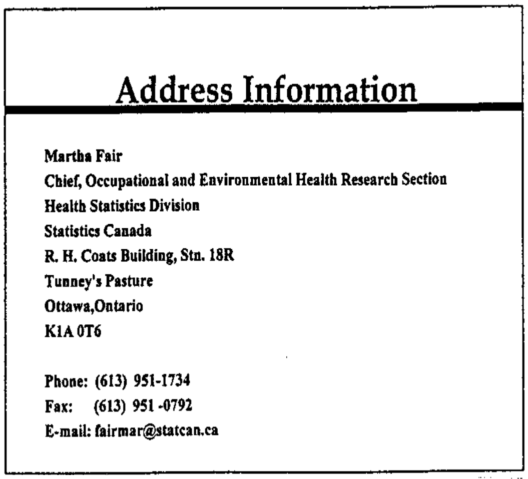
Summary

- Record linkage software development
- Quality of data files
- Uniform classification standards
- Analysis of data
- Analysis of data - incorporating uncertainty due to linkage



Summary

- Be useful
- Expand our horizons
- Ask the right questions
- Know the issues and be aware of priority information needs
- Build the right data and information
- Harmonize concepts and outputs
- Confidentiality protection



Address Information

Martha Fair
Chief, Occupational and Environmental Health Research Section
Health Statistics Division
Statistics Canada
R. H. Coats Building, Stn. 18R
Tunney's Pasture
Ottawa,Ontario
K1A 0T6

Phone: (613) 951-1734
Fax: (613) 951 -0792
E-mail: fairmar@statcan.ca

Appendix A -- Glossary of Terms

There are various terms used in record linkage. Some of these have been defined in: Newcombe, H.B. (1988). *Handbook of Record Linkage Methods for Health and Statistical Studies*, Administration and Business. Oxford, U.K. Oxford University Press, pp. 103-106.

The terms used in that book are as follows:

Blocking. -- The use of sequencing information (e.g., the phonetically coded versions of the surnames) to divide the files into "pockets." Normally, records are only compared with each other where they are from the same "pocket," i.e., have identical blocking information. The purpose is to avoid having to compare the enormous numbers of record pairs that would be generated if every record in the file initiating the searches were allowed to pair with every record in the file being searched.

Denominator. -- This usually refers to the denominator in a FREQUENCY RATIO, i.e., the frequency of a given comparison outcome among UNLINKABLE pairs of records brought together at random. It may be applied also to one of the two components of any ODDS.

Frequency Ratio. -- The frequency of a given comparison outcome among correctly LINKED pairs of records, divided by the corresponding frequency among UNLINKABLE pairs brought together at random. The comparison outcome may be defined in any way, for example as a full agreement, a partial agreement, a more extreme disagreement, or any combination of values from the two records that are being compared. The FREQUENCY RATIO may be specific for the particular value of an identifier when it agrees, or for the value of the agreement portion of an identifier that partially agrees, or it may be non-specific for value.

General Frequency. -- A weighted mean of the frequencies of the various values of an identifier among the individual (i.e., unpaired) records of the file being searched. It is non-specific for value. Value-specific frequencies are also obtained from the same source.

Global Frequency. -- The frequency of a comparison outcome among pairs of records, when that outcome is defined in terms that are non-specific for the value of the identifier. The outcome may be a full agreement, a partial agreement, or a more extreme disagreement. The record pairs may be those of a LINKED file, or they may be UNLINKABLE pairs brought together at random. Only in the special case of the full agreement outcomes are the global and the general frequencies numerically equal, but they always remain conceptually different. The difference is that a global frequency, although value non-specific, always reflects the full definition of the non-agreement portion of that definition. A general frequency cannot do this because it is based on a file of single (i.e., unpaired) records.

Global Frequency Ratio. -- The ratio of the global frequency for a particular comparison outcome among LINKED pairs of records, divided by the corresponding frequency among UNLINKABLE pairs. It is equivalent to the global ODDS. GLOBAL FREQUENCY RATIOS for agreement outcomes and partial agreement outcomes are often subsequently converted to this value-specific counterparts during the linkage process. The conversion is accomplished by means of an adjustment upwards where the agreement portion of the identifier has a rare value, and an adjustment downwards where the value is common.

Linkage. -- In its broadest sense, RECORD LINKAGE is the bringing together of information from two or more records that are believed to relate to the same "entity." For an economic or social study, the "entities" in question might be farms or businesses. For a health study, the "entities" of special interest are usually individual people or families. It is in the latter sense that the word is used throughout this book.

Linked. -- In line with the above definition of "record linkage," LINKED pairs of records are pairs believed to relate to the same individual or family (or other kind of entity). Record pairs brought together and judged not to relate to the same individual or family may be referred as "UNLINKABLE" pairs. For short, the two sorts of pairs are sometimes called "LINKS" and "NON-LINKABLE," respectively. As used here, the term implies that some sort of decision has been reached concerning the likely correctness of the match.

Matched. -- This word is variously used in the literature on record linkage. In this book, however, it is given no special technical meaning and merely implies a pairing of records on the basis of some stated similarity (or dissimilarity). For example, early in a linkage operation, records from the two files being LINKED are normally matched for agreement of the surname code. The resulting pairs may also be called "candidate pairs" for linkage, but this emphasis is most appropriate in the later stages when the numbers of competing pairs have diminished. Pairs of records will frequently be spoken of as "correctly matched," "falsely matched," or "randomly matched."

Numerator. -- This usually refers to the numerator in a FREQUENCY RATIO, i.e., the frequency of a given comparison outcome among pairs of records believed to be correctly LINKED. It may be applied also to one of the two components of any ODDS.

Odds. -- This word is used in its ordinary sense but is applied in a number of situations. As relating to a particular outcome from the comparison of a given identifier it is synonymous with the FREQUENCY RATIO for that outcome. As relating to the accumulated FREQUENCY RATIOS for a given record pair it refers to the overall RELATIVE ODDS. It is also applied to the overall ABSOLUTE ODDS.

Outcome. -- This refers to any outcome or result from the comparison of a particular identifier (or concatenated identifiers) on a pair of records, or the comparison of a particular identifier on one record with a different but logically related identifier on the other. It may be defined in almost any way, for example as an AGREEMENT, a PARTIAL AGREEMENT, a more extreme DISAGREEMENT, any other SIMILARITY or DISSIMILARITY, or the absence of an identifier on one record a s compared with its presence or absence on the other. An outcome may be specific for a particular value of an identifier (e.g., as it appears on the search record) or for any part of that identifier, especially where there is an agreement or partial agreement; it may be non-specific for value; or it may even be specific for a particular king of DISAGREEMENT defined in terms of any pair of values being compared.

Value. -- An identifier (e.g., an initial) may be said to have a number of different "values" (e.g., initial "A," initial "B," and so on). Surnames, given names, and places of birth have many possible values. Other identifiers tend to have fewer values that need to be distinguished from each other.

Weight. -- In the literature, this term has been widely applied to the logarithms of various entities, such as:

- a FREQUENCY RATIO for a specified outcome from the comparison of a given identifier;
- the product of all the FREQUENCY RATIOS for a given record pair;
- the NUMERATOR of a particular FREQUENCY RATIO;
- the DENOMINATOR of a particular FREQUENCY RATIO;
- any estimate of such a numerator or denominator, not obtained directly from a file of matched pairs of records.

The use of the logarithm is merely a convenience when doing the arithmetic; it does no affect the logic except to make it appear more complicated. The term “WEIGHT” has therefore been employed sparingly in this book. Instead, reference has been made directly to the source frequency or FREQUENCY RATIO, or to the estimates of these, wherever possible.